THE REVISION OF THE STANFORD-BINET SCALE

An Analysis of the Standardization Data

BY

QUINN McNEMAR

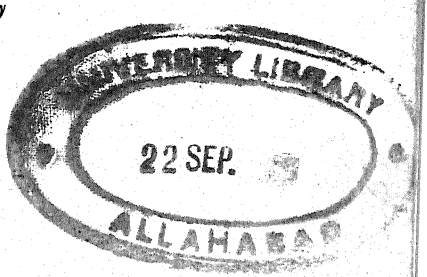
Associate Professor of Psychology
Stanford University

WITH AN
INTRODUCTORY CHAPTER

BY

LEWIS M. TERMAN

Professor of Psychology
Stanford University



HOUGHTON MIFFLIN COMPANY

BOSTON · NEW YORK · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO

PREFACE

The basic data for this volume were collected during the standardization testing for the new Stanford-Binet revision. Although a portion of the material herein is based upon analyses which were essential in the standardization procedure, a large part represents analyses which have been made subsequently. Some of the data pertain to the scale as an instrument, some are concerned with results obtained thereby. At times the chief concern is with the scale as a whole, at times with analyses based on items. It has not been feasible to present more than a statistical summary of the mass of accumulated data; it is hoped that sufficient detail has been given to make the discussion intelligible.

This volume is so closely related to Terman and Merrill's *Measuring Intelligence* that the acknowledgments made therein could be repeated here. In addition, the writer is indebted to the Social Science Research Council of Stanford University for support which made the factor analyses possible, and to the Committee on Psychology and Anthropology of the National Research Council for a grant-in-aid for the study of scatter. The analysis of the data for sex differences was financed in part by funds granted Professor Terman by the Committee on Sex Research of the National Research Council. I am personally indebted to Professors Terman and Merrill for their cooperation and encouragement. Much of the basic work for this volume was done under their direction. Credit is due Dr. Merrill for choosing the items for the nonverbal and memory scales, but she should not be held responsible for my interpretations thereof. Dr. Merrill has also rendered invaluable assistance in assembling the material for Appendix C. Much of the responsibility entailed in the preparation of this volume has been shared by Olga W. McNemar. To Professor Terman I am grateful for his willingness to write the introductory chapter and for many helpful suggestions and criticisms.

Quinn McNemar

CONTENTS

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

Chapter I

THE REVISION PROCEDURES

by

Lewis M. Terman

Dr. McNemar has asked me to prepare an introductory chapter on the purpose of the New Revision, the selection of test items, and the procedures employed in getting the standardization data and constructing the scales. This account can be brief, for the essential facts have been presented in considerable detail in the first three chapters of *Measuring Intelligence*.[1]

Purpose and Character of the Revision

The purpose of the revision was to replace the single Stanford-Binet scale of 1916 by two alternative scales, different in content but functionally equivalent in every way, which would test intelligence more accurately and over a wider range than had hitherto been possible by scales of the Binet type. The 1916 revision, although successful beyond all expectations of the author, had a number of defects that needed to be remedied. It did not extend low enough or high enough, the accuracy of standardization was uneven, and the procedures for giving and scoring the individual tests were not always sufficiently defined. It was intended that the new scales should test as nearly as possible the same aspects of intelligence as did the earlier revision. The fact that the latter had proved its value as a clinical tool and had be-

[1] By L. M. Terman and M. A. Merrill, Boston: Houghton Mifflin Company, 1937.

come the most widely used of psychometric devices was deemed warrant enough for replacing it by something similar and better rather than by something different in kind.

Although no system of psychological tests is ever as good as one would like, the main objections of the revision were reasonably well approximated. The resulting scales are mutually equivalent with respect to difficulty, range, reliability, and validity; they differ from their predecessor chiefly in range, accuracy of standardization, and refinements of procedure. The 1916 Stanford revision contained 90 tests (compared with 54 in the 1908 scale of Binet) and was standardized on 905 subjects in California and Nevada. Form L and Form M of the New Revision have 129 tests each, and were standardized by administration of both forms to 3184 subjects selected so as to sample the white child population of the main geographical areas of the United States.[1] It is not surprising that an undertaking of such huge proportions required several years for its completion. The chronology of the revision was roughly as follows:

The first year was devoted to a search of the literature for test suggestions and to the task of devising new test items. The second and third years were devoted to the preliminary trial of many types of new test items and to experimentation on various procedures for giving and scoring. During the fourth year two experimental forms were made up, detailed directions were prepared for their administration, and seven experienced examiners were put through a course of training in preparation for their field work in gathering the standardization data. During the next two years — the fifth and sixth — the field examiners gave full time to the administration of the trial forms. During the seventh year

[1] At the time of the rescoring in the central office, 12 cases were dropped because of examiner's notations indicating negativism, lack of cooperation, etc.

Miss Mayer and Mrs. Oden checked the scoring of all responses on the 6368 blanks, investigated the effects of alternative scoring procedures, and made up lists of satisfactory and unsatisfactory responses. Two additional years — the eighth and ninth — were required for the Hollerith treatment of the data and for making up successive trial revisions until final scales were derived which were mutually equivalent and so standardized for difficulty as to yield mean I.Q.'s of approximately 100 at all age levels.

Selection of Test Items

During the fifteen years prior to the beginning of work on the New Revision a large amount of information had accumulated on the behavior of various types of mental tests. Thousands of correlation coefficients had been published which threw new light on the interrelationships of tested abilities and on the relative merits of specific tests as measures of general intelligence. Many long-debated questions had been answered by experience. Some tests formerly regarded as mere tests of erudition had been found to be highly saturated with a general factor; others formerly looked upon with favor had been found inferior or near-worthless. When the present undertaking was launched it was possible to forecast with considerable assurance whether a new test of a given type would correlate highly, moderately, or only slightly with a battery of tests like the 1916 Stanford-Binet. There would still be surprises, but the period of blind groping was over.

In preparation for the New Revision hundreds of test items were devised and assembled for preliminary tryout. The criterion for inclusion in the preliminary series was whether the test would probably correlate well with the best current batteries of intelligence tests. Experimental tryout of the new test items followed. This

involved their application to groups of subjects who had been tested by the 1916 Stanford-Binet. Test items which showed a rapid rise in per cents passing at successive mental-age levels were tentatively selected for the final tryout if otherwise satisfactory. Items that did not satisfy this criterion were at once discarded.

Mental-age groups were not available for the preliminary tryout of items in the lowest age ranges, and here it was necessary to use chronological-age groups. At this level, where adjacent ages overlap much less in mental ability than they do at the higher levels, the per cents passing at successive ages must necessarily increase rapidly if the test is a good measure of mental growth, although of course it does not follow that a steep age curve guarantees validity. This lack of mental-age groups for use in the preliminary tryout with young children did not appreciably affect the ultimate choice of test items, for in the final tryout the items which did not show a reasonably good correlation with the composite score on the two scales were eliminated. That is, the ultimate criterion of validity was correlation with mental age or its equivalent in point score on the composite of the two scales. In no case was per cent passing by chronological age the final criterion of validity. It is necessary to emphasize this fact because of misstatements that have been made by some of the critics of the age-scale method.

Although a reasonable degree of validity (as above defined) was a necessary requirement for the retention of tests, it would not have been practicable to base the selection of tests upon this criterion alone. Other factors that had to be considered included time for administration, ease and objectivity of scoring, appeal to the subject, the desirability of variety in the sampling of mental processes, and the available supply of tests suitable at any particular level.

As for time, it was desired that the entire examination by the revised scale should ordinarily not extend

much beyond 75 minutes for older subjects or beyond 50 minutes for the younger. It was thus necessary to choose between many brief tasks or fewer long ones. The first of these alternatives has two distinct advantages: it permits a wider sampling of mental behavior and it makes greater appeal to the child's interest. Moreover, it is probably favorable to the validity of the scale as a whole. Consider, for example, a test that requires 4 minutes and correlates .45 with total score, and two others that require only 2 minutes each and yield correlations of only .40 with total score. In such a case it is quite likely that the two briefer tests combined will have a higher correlation with total score than will the single 4-minute test. In general it was the policy of the authors to choose the briefer of two tests when other things were not too unequal.

Although the final trial series contained about 50 per cent more tests than it was planned to use, so many eliminations were necessary that at certain levels there was a shortage of good items and some had to be retained that were of relatively low validity or otherwise not entirely satisfactory. It may interest the reader to compare the first-factor loadings of the different tests as recorded in Tables 29-42 of Chapter IX. Dr. McNemar has made up lists of tests having highest or lowest first-factor loadings at the lower, middle, and higher age levels (124-137).

The scales as published contain 129 tests each. In the final trial series Form L contained 209 tests and Form M 199. As it turned out, this seemingly liberal allowance for rejections was somewhat less than it should have been. The outcome would have been more satisfactory if a larger margin had been provided, say an excess of 100 per cent over the number of items that would be needed. The greater the amount of preliminary sifting the less, of course, will be the need for excess items in the final trial forms.

The Binet scale has often been criticized because of its great variety of brief, disconnected tests — a 'motley array,' as Spearman scornfully refers to them. According to some critics, if there is any large general factor measured by such tests it is purely accidental. They contend that the logical way to proceed is to devise a few series of tests, each series containing many items of a given kind and so, presumably, measuring thoroughly a given aspect of intelligence. The method they recommend has its advantage in group testing in that it simplifies administration procedures, but no test of this type has ever been devised that rivals the Binet test for clinical use with children. The latter, for all its 'motley array' of variegated tasks, not only is more interesting but also affords a better measure of all-round intellectual development than any of the substitutes that have been suggested. Binet's abandonment of the attempt to test the intellectual 'faculties' as such was his outstanding contribution to psychometrics.

The Standardization Group

Various types of communities in eleven states were sampled to secure the 3184 subjects composing the standardization group. There were approximately 100 subjects at each half-year interval from 1-1/2 to 5-1/2 years, 200 at each age from 6 to 14, and 100 at each age from 15 to 18. Every age group was equally divided between the sexes. All subjects were of the white race and American born.

Elaborate precautions were taken to make the sampling as representative of the entire population as circumstances permitted. These have been fully described in *Measuring Intelligence* (pages 12 - 21) and need not be repeated here. They included such considerations as choice of geographical localities, choice of urban, sub-

urban, and rural communities, selection of schools within a given community, and methods of obtaining random age samplings wherever testing was done. It was the age samplings below 7 and above 14 that gave the most trouble. During the progress of the field work comparisons were made between the cumulative sampling and various census classifications of the general population, so that low spots might be filled as the testing proceeded. It was not possible, however, to provide an adequate sampling of rural subjects, because of the labor and expense involved in moving from one small school to another. This was less serious than might be supposed, since it was possible on the basis of found differences in mean I.Q. between rural and urban subjects to estimate and allow for the error caused by inadequate rural sampling.

One of the procedures followed deserves special mention, namely, limitation of the group to subjects who were within one month of a birthday (or half-year birthday for subjects at ages 1-1/2 to 5-1/2). Besides giving relatively homogeneous age groups, this procedure provided a more random sampling than would have been secured by the usual method. The effect, of course, was to multiply the spread of sampling by 6 from age 6 upward, and by 3 from age 6 downward, thus causing less to depend upon the choice of a given school or community.

Apart from the inadequate number of rural subjects it is difficult to see how the sampling could have been much better than it was from ages 7 to 14. Above and below these levels, despite all the precautions taken, it cannot be guaranteed. Below age 4 and above 14 the sampling is almost certainly skewed in the direction of too high I.Q.'s, a probability that was taken into account when the final forms of the test were made up.

Testing Procedures

However large and representative the standardization group might be, the value of the obtained data

would depend ultimately on the expertness with which testing was done. As a result of long experience in the training of examiners the authors of the revision were aware of the amount and kinds of error that may be introduced by careless test procedures. The precautions taken included (1) the selection of experienced examiners who were known to be expert and dependable, (2) careful training of the examiners in the new procedures, (3) wide-range testing, (4) provision for quiet and seclusion during the examination, (5) the administration of both scales to each subject within a period of a week or less, (6) frequent instructions to examiners on methods of meeting new difficulties encountered in administration of scoring, (7) the rescoring of all blanks after the field work was completed, (8) the checking of all M.A. and I.Q. computations, (9) allowance for practice effect resulting from the first scale administrated, and (10) the Hollerith recording and treatment of scores.

It can perhaps be said that no other large group of subjects was ever Binet-tested with so much attention to accuracy and thoroughness, and that the resulting scores, especially the composite scores of the two scales, are the most dependable that have ever been reported.

In view of the fact that the administration of a mental test is in reality a psychological experiment, the need for carefully defined procedures to insure objectivity would seem too obvious to call for special comment. However, strange as it may seem, there are still clinical psychologists who prefer 'a flexible test that can be adapted to the individual,' one that 'will be custom-made to fit each subject.'[1] Needless to say, the progress of psychometrics has consisted largely in escape from the chaos of subjectivity resulting from the impromptu procedures advocated by the author just quoted.

[1] G. H. Kent, "Suggestions for the Next Revision of the Binet-Simon Scale," *Psychol. Rec.*, 1937, 409-432.

THE REVISION PROCEDURES

Age Placement of the Tests

Age placement of the tests in the trial forms was provisional and was intended chiefly to insure that a subject's range of possible success and failure would be fully covered without waste of time by unnecessary testing. It was only after the data were in that the relocating of tests for the final scales was undertaken.

The primary objective was an arrangement of the tests that would yield mean I.Q.'s as near as possible to 100 at all age levels. This had to be done empirically, for there seems to be no possibility of a mathematical solution when there are so many test items and no two of them behave exactly alike with respect to curves of per cents passing them. The plan which some have advocated whereby all tests would be located at the age where 50 per cent of unselected subjects pass them simply does not work, as it yields mental ages that are much too high in the lower range and much too low in the upper range. For a scale of the Binet type there is no one 'correct' per cent for locating all the tests. The fact that adjacent mental ages become progressively closer together from the lower to the upper ranges, with the scatter of an individual's performance increasing correspondingly, means that tests located correctly at a lower age will show a higher per cent of at-age passes than will a test correctly placed at a higher age. Moreover, the correct placement of tests for a particular age depends in part on the tests in the preceding and succeeding ages. For example, the correct per cent of at-age passes for tests in year XII depends partly on whether there are tests at years XI and XIII, as in the New Revision, or none, as in the original Stanford-Binet.

The relocation of tests was done first for Form L. A tentative rearrangement was made and the resulting I.Q. distributions by age were examined. A second revision followed and the new I.Q. distributions were examined. Several revisions of this kind were necessary

before the standardization of Form L was regarded as satisfactory. When the task had been completed for Form L it was possible to standardize Form M at once by matching its tests age for age with the tests of Form L on curves of per cents passing.

One of the most serious problems was caused by the lack of enough test items of exactly the right difficulty at certain levels. This could sometimes be handled by rescoring a test on a different standard of pass or failure, and thus making it easier or harder as the situation demanded. In a few cases, however, relatively inferior tests had to be retained, and there are 16 test items that were included in both scales.[1]

The placement of tests at the three adult levels, where the mental-age scores are fictitious units, was governed by the same requirement as at other levels, namely, that the resulting mean I.Q. in the standardization sample should approximate 100 for the subjects at each chronological age. In computing I.Q.'s of older subjects, chronological age was dropped off gradually instead of all at once. Four months were dropped in the 14th year, four in the 15th, and four in the 16th. That is, maximum C.A. used in the divisor is now 15. The arrangement of tests at the upper levels yields I.Q. distributions very similar to those at lower levels; this was not true of the 1916 revision.

At all scale levels effort was made to have the tests within a given age group of equal difficulty. That this could only be approximated is not serious; minor inequalities have no effect other than to increase somewhat the danger that a test may be omitted that should have been given. It should be understood, moreover, that the order of difficulty of the tests for our standardization

[1] The duplicated items are scattered over a considerable part of the scales and are not numerous enough in the range over which a single subject would be tested to affect appreciably the correlation between the two scales.

group is not necessarily that which will be found for special groups such as older adults, psychopathic subjects, Negroes, Indians, etc.

Age Scales versus Point Scales

The choice between age scale and point scale is not necessarily a choice between M.A. and I.Q. scores on the one hand and point scores on the other. Any point scale designed for children will of course be provided with age norms, and these age norms are nothing more nor less than mental-age scores. Moreover, the inevitable comparison of M.A. and C.A. brings us back to I.Q. scores. However strongly the makers of point scales condemn the M.A. and I.Q. as measuring units, the users of such tests always revert to them, at least as long as they are dealing with children.

Some point-scale authors are so allergic to Binet I.Q.'s that in trying to avoid them they fall into amusing statistical pitfalls. For example, Wechsler[1] proposes that instead of expressing I.Q. as $\frac{\text{M.A.}}{\text{C.A.}}$ it should be expressed as $\frac{\text{attained or actual (point) score}}{\text{expected mean (point) score for age}}$. This method leads to interesting results when the point scale has an arbitrary zero point, as nearly all of them do. By arbitrary zero point is meant one that represents something higher than zero ability. Such a scale is analogous to a measuring-stick on which 24 inches is called zero, 25 inches is called 1 inch, 26 inches is called 2 inches, etc. When the height of an infant is measured by such and turns out to be 1 inch (really 25 inches), whereas the norm for his age is 2 inches (really 26 inches), the infant's Height Q. is exactly 50. If measured height

[1] David Wechsler, *The Measurement of Adult Intelligence*. Baltimore: Williams and Wilkins Company, 1939; see p. 25.

had been .2 inch (really 24.2 inches), the H.Q. would have been 10. That is, height deviations within the small range of .8 inch from the norm would yield H.Q.'s from 10 to 100. Vocabulary quotients exactly analogous to these absurd height quotients have actually been used.

As we have noted above, however, point scores can be converted into M.A.'s and I.Q.'s which are comparable with the M.A.'s and I.Q.'s of an age scale of the Binet type. In view of the fact that a point scale is more easily standardized, since it is not necessary to juggle the tests about until they have been properly assigned to age groups, the question may be raised why the authors of the New Revision went to so much unnecessary labor to produce age scales. The answer is that the age arrangement is preferred by a majority of examiners because it enables them to follow more intelligently the test-by-test progress of a subject during the course of the examination. The authors believe that this advantage of the age scale warranted its extra cost in time and labor.

For a mathematical discussion of units of measurement the reader is referred to Dr. McNemar's treatment of the subject in Chapter XI.

Outcome of the Standardization

No extended discussion of the results of the standardization is here necessary since the problem is treated from various angles in the chapters that follow. A few points may perhaps be emphasized without duplicating unduly the material of other chapters.

Accuracy of Standardization. - The authors purposely adjusted the scales so that mean I.Q.'s of the standardization group would be slightly above 100. The main justification for this was the inadequate sampling of rural subjects, previously referred to. Additional allowance in this direction was made for ages below 4 and

above 15 because of reasons for believing that the sample tested was definitely superior at these levels. The means for the two scales run closely parallel and the smoothed means for the composite of L and M (*Measuring Intelligence*, page 36) give the truest picture of the accuracy with which the standardization fits the sample tested. The greatest difference between any two means in the entire range is 4.3 I.Q. points. From ages 4 to 15 inclusive the greatest difference is 3.2 points. Of the fourteen means in this range, ten differ from 100 I.Q. by 0.0 to 1.6 points and four by 2.0 to 3.0 points.

Reliability. - It would appear from the data presented in Chapter VI that a degree of reliability has been attained that would be difficult to exceed without extending the examination beyond practicable time limits. The reliabilities have been expressed in terms of both σ_e (standard error of measurement) and the equivalent reliability coefficients. An important fact brought out by Dr. McNemar is that σ_e varies directly with size of I.Q. It may be reassuring to clinicians who work largely with backward subjects to know that the New Revision is particularly reliable at low I.Q. levels.

Attention may be called to the fact that since all subjects were given both the L and M scales, and since the tests were administered by expert examiners, the data give an exceptionally accurate picture of the reliability factor.

Validity. - The somewhat futile war of words regarding the validity of this or that intelligence test has died down as a result of the growing custom of defining validity in operational terms. A test tests what it tests, and the nature of this 'what' only becomes clear as the test is used and the results checked. Here it is perhaps sufficient to note that Forms L and M correlate with the 1916 revision about as highly as their respective reliabilities permit. The new scales test whatever the old one tested, but with somewhat greater accuracy.

The fourteen factor analyses which Dr. McNemar has made of the tests at successive levels throw some light, despite their avowed limitations, upon factors measured. In particular they indicate that at no level do the tests measure a medley of factors as some have believed; everywhere there is one factor that stands out clearly with only occasional and none too reliable evidence for a second or third factor. Furthermore, by the ingenious method of having the adjacent analyses overlap in respect of test items included, Dr. McNemar has demonstrated that sudden changes in the nature of the primary factor do not occur from level to level. His data in fact indicate, although they do not fully demonstrate, that the primary factor is the same at widely separated levels. This tentative conclusion could be checked by retesting a large group of subjects over a period of ten or fifteen years.

* * * * * * * * * * * * * *

This report on the standardization data from which the 1937 Stanford-Binet scales were derived has long been overdue. Its appearance at this time has been made possible by the fact that Dr. McNemar, on the urgent request of the New Revision authors, consented to write the several chapters for which they had originally been scheduled. The delay in publication, although regrettable, is offset by the high professional quality of the job Dr. McNemar has done.

Chapter II

ON THE DISTRIBUTION OF I.Q.'S

Since the introduction of Pearson's system of frequency curves, a large amount of energy has been expended in graduating frequency distributions. Many early biometric studies had as their sole objective the exact mathematical specification of the distribution of some characteristic. These studies definitely showed that the types of distribution are many, and that the normal curve is far from being the rule. It was only natural for psychologists to be influenced by the early work of the biometricians, but unlike their predecessors, the psychologists seemed to find normal distributions more often than other types. This led to the assumption that the normal curve was the ideal, and that exceptions thereto should be explained. It also led some to speculate, via the binomial expansion, as to the nature of the constituent elements of mental life.

That the shape of the distribution for any psychological trait is partly dependent upon the units of measurement employed is so axiomatic that one wonders why it should have ever been assumed that deductions could be made concerning whether the underlying, indirectly measured, trait is normally distributed. Furthermore, the ease with which the shape of a distribution can be altered by a change in test difficulty should also have served as a warning to those who were out to demonstrate the normal law for psychological traits. It is our contention, certainly not original with us, that nothing can be inferred from the distribution of measured psychological traits with regard to the shape of the distribution which would result if we ever found a psychometric of truly equal units.

The most ambitious attempt to show that intelli-

gence is normally distributed is to be found in Appendix III of Thorndike's *Measurement of Intelligence*. Therein are given distributions for single tests which vary to an unspecified degree from that expected on the basis of the normal law. Composite distributions, secured by averaging the frequencies for single tests, were so closely fitted by the normal curve as to leave no doubt in one's mind about the goodness of fit - too good to be above suspicion. His very high chi square probability figures of order .99 to .999999 simply indicate either statistical misapplication or an undue restraint on the operation of chance. It happens that the process of averaging frequencies does interfere with the type of chance discrepancies which might be tested by chi square; hence we do have a faulty use of statistical method. Aside from the unacceptable chi square probabilities, the fact remains that the histograms based on composites at least appear to be normal. It seems safe to assume that the makers of the several single tests or scales were guided more or less by the practical advantage inherent in a test which yields a normal distribution, and it seems likely that various factors are involved in the lack of normality for the several single curves. When ten or eleven such curves are averaged, one might expect a balancing of these factors; so if the resulting composite turns out to be normal, it merely demonstrates that the averaging of frequencies based on several tests involving arbitrary scoring units may possibly lead to a normal curve. This does not prove that intelligence is distributed normally, unless one is willing to assume that such averaging of frequencies (ordinates) has somehow ironed out the inequalities of the arbitrary units along the abscissae. One would need an abundancy of faith in the law of averages plus a belief in statistical miracles before accepting such an argument.

Now Thorndike has faced the problem of inequality, or the unknown equivalence, of units of measurement. His Chapter VIII is devoted to the proposition that intel-

ligence is distributed normally provided the original score units are first transmuted into a scale of 'truly equal units.' Since this transmutation is accomplished by means of normal curve functions, one need not be surprised to find that the distributions, so adjusted, seem to substantiate the normal hypothesis. This is especially true (and to be expected) when composites, including the total group upon which the scaling was based, are considered. This entire procedure is another example of man trying to lift himself by tugging at his own boot straps. It is our opinion that we have no exact knowledge about the distribution of intellect, and that Thorndike has only demonstrated the obvious: a normal curve can be produced by assuming it in advance.

We have already indicated another way of securing a normal curve, namely, by selecting items which are of medium difficulty for a given group. A test so constructed will yield distributions closely approximating the normal form for groups which are similar in level of ability and homogeneity to the starting group, but will likely yield skewed distributions for groups of other ability levels. There can be no objection to this procedure provided no claims are made regarding the normality of the underlying trait. In fact, it is convenient to have a normal distribution for a trait as measured since the statistical aspects of the Gaussian curve are relatively simple and generally known. In revising the Binet scale, no attempt was made, as erroneously claimed by some, to secure a normal distribution of the resulting I.Q.'s. Items were chosen in such a manner that the average M.A. for an age group coincided with their C.A. The per cent passing with age curves for the items were ogival, but not necessarily normal ogives. The one thing about the scales which would tend to produce a normal curve of distribution is the fact that each age group is tested by items which cover the entire range of difficulty. This means that in testing a group of a given age, the varying difficulty of all the items attempted is such that it can be

said that on the average the items are of medium difficulty for that particular group. Consequently one might very well expect the M.A.'s and I.Q.'s for a single age, and the I.Q.'s for several ages combined, to approximate normality.

Frequency distributions for Forms L and M separately for three age groupings and for all ages combined are presented in Tables 1 and 2. At the bottom of each table will be found the N's, means, S.D.'s, measures of skewness and kurtosis (based on moments) and their standard errors,[1] and the chi square probabilities for obtaining as large discrepancies from the best-fitting normal curves. (The grouping of end intervals for the computation of chi square is indicated by braces.) When the values for g_1 (skewness) and g_2 (kurtosis) do not depart significantly from zero, it can be said that the hypothesis of normality is tenable provided the chi square probability is not less than, say, .01. For all ages combined, the observed distributions, reduced to percentage frequencies, and the superimposed normal curves are pictured in Figures 1 and 2.

From the data and constants presented in Tables 1 and 2, it seems safe to set forth the following brief summary: For ages 2-1/2 to 5-1/2, both forms, the distributions are reasonably symmetrical but more peaked than normal distributions. For ages 6 to 13, again both forms, the skewness and kurtosis are not significantly different from normal, but the irregularities are highly significant as indicated by the small chi square probability figures. For ages 14 to 18, it can be said that both forms yield distributions which, for the sample at hand are not significantly different from the normal form. For ages 2-1/2 to 18 combined, the distributions are lacking in skewness, but are more peaked than expected on the basis of the normal curve.

[1] Cf. pp. 78-79 in R. A. Fisher's *Statistical Methods for Research Workers.* (6th Ed.) Edinburgh; Oliver and Boyd, 1936.

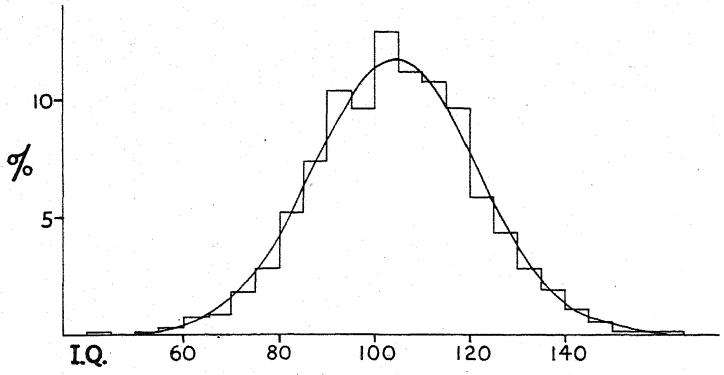


Fig. 1. Form L I.Q. distribution and best-fitting normal curve, ages $2\frac{1}{2}$ to 18

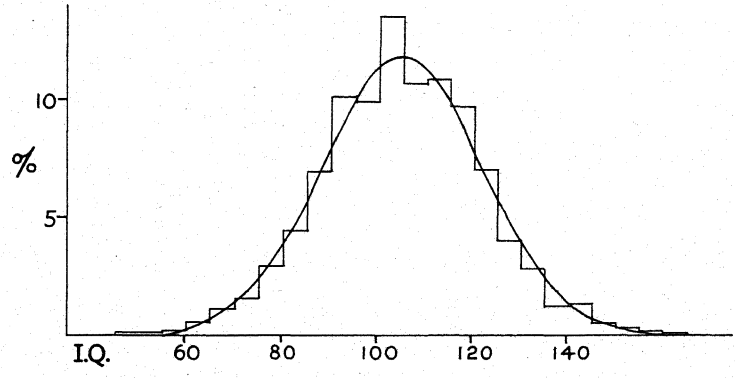


Fig. 2. Form M I.Q. distribution and best-fitting normal curve, ages $2\frac{1}{2}$ to 18

The exhibited departures from normality are scarcely large enough to disturb practical interpretations of I.Q.'s as though normally distributed. The degrees of skewness, kurtosis, and irregularity of the distributions are certainly not so great as to invalidate the use of sampling and correlational techniques in researches involving the Stanford-Binet I.Q. as a variable.

We draw no conclusions from these data concerning the distribution of intellect. As measured by the New Stanford Revision of the Binet tests, I.Q.'s are approximately normal in distribution, but we make no claims for the equality of the units involved.

Incidentally, the fact that the means in Tables 1 and 2 are above 100 should not lead the reader to the erroneous conclusion that the average I.Q. for the population now exceeds 100. The excess here observed is in the proper direction to allow for known bias in our age samplings. When an adjustment is made for bias in occupational status, the age means approach nearer 100, and a further adjustment for inadequate rural representation would tend to bring the values still closer to 100.

TABLE 1

FREQUENCY DISTRIBUTIONS AND CONSTANTS FOR FORM L I.Q.'S

	2-1/2 - 5-1/2	6 - 13	14 - 18	2-1/2 - 18
170		1		1
165	1	0		1
160	3	1 } 12		4 } 13
155	2 } 13	1		3
150	2	2		4
145	5	7	2	14
140	6	22	4 } 20	32
135	12	31	14	57
130	29	42	12	83
125	35	65	28	128
120	47	75	51	173
115	87	145	53	285
110	88	163	68	319
105	93	184	52	329
100	100	223	58	381
95	47	150	87	284
90	60	185	61	306
85	44	125	47	216
80	32	90	31	153
75	14	43	26	83
70	6	33	13	52
65	4	19	1	24
60	3	11	8	22
55	3	2	2	7
50	2 } 15	2 } 16	0 } 12	4
45	1	0	0	1 } 16
40	1	1	1	3
35	1			1
N	728	1623	619	2970
M	106.58	103.22	103.03	104.00
S.D.	17.32	16.83	16.89	17.03
g_1	-.131	.133	-.090	.028
σ_{g_1}	.091	.061	.098	.045
g_2	1.117	.209	-.150	.346
σ_{g_2}	.182	.122	.196	.090
x^2 P	.006	.003	.06	.03

TABLE 2

FREQUENCY DISTRIBUTIONS AND CONSTANTS FOR FORM M I.Q.'S

	2-1/2 - 5-1/2	6 - 13	14 - 18	2-1/2 - 18
165		1		1
160	2	0		2 } 16
155	2	3 } 16		5
150	3 } 19	5		8
145	4	7	5 } 11	16
140	8	25	6	39
135	9	19	8	36
130	24	38	22	84
125	35	63	23	121
120	62	107	40	209
115	75	151	62	288
110	92	166	64	322
105	73	188	56	317
100	112	212	78	402
95	67	160	66	293
90	59	182	58	299
85	35	117	54	206
80	28	79	25	132
75	18	43	24	85
70	10	25	10	45
65	3	21	9	33
60	1	7	7	15
55	2	2	1 } 18	5
50	1 } 10	1 } 11	0	2
45	2	0	1	3 } 12
40	0	1		1
35	1			1
N	728	1623	619	2970
M	106.42	103.96	103.32	104.43
S.D.	16.72	16.55	17.11	16.75
g_1	-.109	.125	-.050	.029
σ_{g_1}	.091	.061	.098	.045
g_2	.925	.218	-.110	.298
σ_{g_2}	.082	.122	.196	.090
χ^2 P	.09	.009	.37	.005

Chapter III

ANALYSIS BY AGE-GRADE

In the previous volume, *Measuring Intelligence*, means and standard deviations for each of the twenty age groups were given for Forms L and M separately. It is the purpose of this chapter to present the data in terms of I.Q. and M.A. for grade and age-grade groupings. We shall be concerned only with those subjects of ages 6 to 18 who were in grades 1 to 12. Since the two forms are so similar that parallel analyses would become highly repetitious, and since scores based on an average of the two forms are more reliable, L-M composite[1] M.A.'s and I.Q.'s are utilized here.

The age-grade distribution of the subjects who were in grades 1 to 12 and of ages 6 to 18 is set forth in Table 3. At this place it should be recalled that the selection of subjects, particularly for ages 6 to 14, was such that no selective factors within a school were operative - all children in a given school who were within one month of a birthday were utilized regardless of their grade location. The schools chosen were of average social status for their communities, which had been selected so as to yield representative samplings. The fact that different

[1]This composite is actually based on penultimate scoring. Minor changes in final scoring procedure for the final normative data resulted in slight changes in the I.Q.'s of a large proportion of the subjects, but these changes were such that shifts occurring in means and standard deviations were negligible as judged from data on age groups 8 to 15 for which comparisons could easily be made. Since the discrepancies in the means and S.D.'s for these eight groups were .5 or less, it would seem permissible to use the penultimate scores in the analyses for age-grade and also for occupational status. The penultimate scores are not being used by preference but because the final scores were not on the particular Hollerith cards which contained grade and occupational information.

communities and schools are here involved does place one limitation on an analysis of intelligence by age-grade, namely, the likely differences in promotion practices in the various school systems. The frequencies in Table 3 cannot be thought of as representing adequately any one community, nor can one be sure of their generality. It is of interest, however, to note that the figures in Table 3 for the elementary grades show essentially the same features as the age-grade distribution of 500,000 California children in 1921-1922.[1]

If one examines Table 3 it will be seen that the 6-year-old subjects were mostly in grade 1 (50 were in kindergarten, and 22 not in school), and that more of the 7-year-olds were in grade 1 than in grade 2. Thence up the diagonal the modal frequencies occur at grade 2, age 8; grade 3, age 9, etc. About these modal age-grade groups the distributions for constant age and for constant grade tend to be skewed. The extremes of retardation are greater than the extremes of acceleration. Furthermore, if the modal values be regarded as representing normal school progress, the number accelerated by one grade exceeds the number retarded by one grade.

Table 4 contains the means for L-M composite I.Q.'s by age-grade, by grade, and by age (the means by age, it should be noted, are for school cases only). There would seem to be a slight relationship between grade location and mean I.Q. (see right-hand margin), even though I.Q. is not related to age except for the slightly higher values for ages 17 and 18 (see bottom margin). The striking, though not surprising, facts in Table 4 are to be observed by following up the modal age-grade diagonal, beginning with the cell for age 7 and grade 1. The means for the modal groups are from 0 to 4 points below the respective age averages. This fact, coupled with the data of Table 3, which show that the frequency for the

[1] See T. L. Kelley, "Ridge-Route Norms," *Harvard Educ. Rev.*, 1940, 10, 309-314.

next grade above the modal grade for a given age is fairly large compared to the frequency in the modal grade, would suggest that the modal age-grade groups do not really represent normal progress pupils in the sense of an average. It would seem likely that modal values based on individuals who were within one month of 7-1/2, 8-1/2, etc. years of age, rather than 7, 8, etc., would provide modal groups which could be taken as a more exact index of normal progress through school.

Turning again to Table 4, it will be observed that those who are accelerated by one grade (we are here accepting the modal age-grade as the norm) tend in general to average about 11 I.Q. points above the normal progress groups, while those who are retarded by one grade are about 11 points lower than the normal. Those cases who are accelerated by two grades and those retarded by two grades deviate about 22 points from the normals, while those children (too few to justify reporting means) whose acceleration or retardation is greater than two grades tend to possess I.Q.'s still farther from those of the modal groups. It is thus seen that school progress for individuals of a given age is definitely related to, if not dependent upon, the intelligence factor. We hasten to point out, however, that we have been dealing with averages and that our figures do not indicate any greater predictability for individuals than has heretofore been possible. The standard deviations (see Table 5) for age-grade I.Q. distributions indicate definitely that age-grade location will not permit a very accurate prediction of an individual's I.Q., and from this we know that the reverse prediction will be far from perfect.

The standard deviations in Table 5 show no particular trend except that, as regards I.Q., an age-grade group is more homogeneous than either an age or a grade group. Considered along with the means of Table 4, they provide an indication of the amount of overlapping between the groups of dissimilar school acceleration or retardation.

ANALYSIS BY AGE-GRADE

Our discussion of the results in Table 4 have so far been chiefly concerned with comparisons about the diagonal in a vertical direction, i.e. with the I.Q.'s of those of a given age but with grade location varying. Let us now re-examine the table, with particular attention to the horizontal rows. For a given grade, age varies, and the mean I.Q. varies inversely with age - the younger in a grade have I.Q.'s above average, while the older have lower I.Q.'s. It does not follow from this that a grade group is more heterogeneous than an age group as regards the I.Q. index of brightness. The standard deviations along the right-hand and bottom margins of Table 5 show that grade groups do not differ appreciably or systematically from age groups in I.Q. variability. Nor does the presence of younger bright children and older dull children in a given grade indicate, ipso facto, that the older dull are handicapped in their competition with the younger and brighter individuals.

Actually, when we turn to Table 6, wherein will be found the mental age means by age, by grade, and by age-grade, we see that the mental maturity of the younger is only slightly greater than that of the older children in the same grade. This means that the acceleration and retardation of our subjects are such that the presence of individuals of varying age in a grade contributes little to the variability of mental age within a grade. As a matter of fact, the acceleration and retardation have in reality operated so as to produce grade groups which are slightly more homogeneous in regard to mental age than would occur in case promotions were made automatically by age only. This is clearly evident from Table 7; the S.D.'s for M.A.'s by grade (right-hand margin) are smaller than the corresponding values for age groups, except for the first two or three grades which are less affected by acceleration and retardation.

Thus it is seen that promotion practices have not materially decreased the wide variation in mental level within a grade from that which would result if promo-

tions were made entirely on the basis of age. The extent of such variation is summarized in Table 8, which shows for each grade the M.A. distance separating the highest sixth from the lowest sixth and the highest 2 per cent from the lowest 2 per cent. It will be noted that the gap between highest sixth and lowest sixth starts at 1.8 M.A. years in the first grade, increases to 2.8 in the fifth grade, and to 4.0 in the eighth grade. The gap between the highest 2 per cent and the lowest 2 per cent starts at 3.6 M.A. years in the first grade, increases to 5.6 years in the fifth and to 8.0 years in the eighth. Beyond the eighth grade, there is little tendency for increase in variation.

In view of the relatively large N's on which the above data are based, the representative nature of the sampling, and the accuracy of the mental-age measure (composite of Forms L and M), there can be no doubt that Table 8 gives an essentially true picture of the results of grading practice in the schools of this country during the 1920's. The facts speak for themselves. A teacher confronted with the necessity of teaching en masse third-grade children ranging in mental age from 6 to 12 years, or sixth-grade children of mental ages from 8 to 16 years, must indeed be versatile if she is to provide learning situations appropriate for all. It is interesting that the widespread use of group mental tests and standardized achievement tests for fifteen years prior to the collection of these data seem to have had very little effect upon the heterogeneity of mental ages in a given school grade. The situation revealed in Table 8 in fact parallels rather closely that found by Terman and his associates between 1915 and 1920.[1]

[1]See especially the following references:
Virgil Dickson, *Mental Tests and the Classroom Teacher.* Yonkers, New York: World Book Company, 1923. L. M. Terman, *The Intelligence of School Children.* Boston: Houghton Mifflin Company, 1919. L. M. Terman et al., *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence.* Baltimore: Warwick and York Inc., 1917.

ANALYSIS BY AGE-GRADE

The fact of individual differences and the question as to what the schools should do about them have provoked far more discussion than research. Few would deny that it is the responsibility of the school to adjust to the developmental needs of the child, but there would seem to be little unanimity as to what type of adjustment is best from the child's standpoint. It is not our purpose to argue here the pros and cons of the various provisions which have been made for the bright and dull. A few observations, however, may be in order. We agree with those who believe that retardation does not solve the problem of the sub-average. Their all-round development might be served better by regular promotions with extensive adjustment of content and activity. That marked acceleration of the above-average may not be the best solution is indicated by the judgments of a large proportion of the Terman gifted group. Some 80 per cent now, as adults, believe that rapid school acceleration was harmful for them. Again it appears that adjustments in the provisions for learning should be made without too much departure from normal grade promotions. Ability grouping within a grade would seem to be an acceptable practice. Those who argue that such a plan tends to stigmatize the dull ignore the likely fact that by the time children are mature enough to appreciate the meaning of such groupings they are also discerning enough to spot the dullard and attach their own classification label to him.

TABLE 3

AGE-GRADE DISTRIBUTION OF SUBJECTS - AGES 6 TO 18, GRADES 1 TO 12

	Age													
	6	7	8	9	10	11	12	13	14	15	16	17	18	6-18
Grade														
12											4	24	39	67
11										5	25	36	29	95
10									5	28	34	22	7	96
9								9	39	40	19	7	2	116
8							5	59	103	25	10			202
7						3	64	90	37	6	1	1		202
6					3	72	88	34	12	3				212
5				2	72	95	30	8	2		1			210
4				83	98	27	12	3	2					225
3		1	65	95	25	6	2							194
2	1	90	119	22	3	1	1							237
1	130	99	19	2										250
1-12	131	190	203	204	201	204	202	203	200	107	94	90	77	2106

TABLE 4

MEAN I.Q. BY AGE AND GRADE

(Means not given where N less than 5; but marginal values include all cases)

Grade	Age 6	7	8	9	10	11	12	13	14	15	16	17	18	C-18
12												113	113	114.8
11										129*	115	108	105	110.2
10									124*	116	102	103	98*	107.0
9								124*	111	103	98	101*		105.7
8							143*	114	103	90	84*			105.1
7							114	102	92	81*				103.5
6						117	102	94	82					104.4
5					112	101	92	80*						102.6
4				111	103	87	74							101.4
3			112	103	87	75*								102.6
2		108	100	86						*N = 5-10				101.7
1	104	97	89											100.0
1-12	103.9	102.9	103.0	104.2	103.9	104.2	103.0	103.9	101.2	102.6	103.3	107.0	108.2	103.69

TABLE 5

STANDARD DEVIATIONS OF I.Q.'S FOR AGE-GRADE DISTRIBUTIONS
(Not given where N is less than 5; marginal values include all cases)

	Age													
	6	7	8	9	10	11	12	13	14	15	16	17	18	6-18
Grade														
12												9	16	14.3
11										13*	13	13	14	14.4
10									9*	12	11	15	13*	14.5
9								16*	15	15	14	11*		16.8
8							8*	15	13	14	13*			17.5
7							15	14	13	7*				16.6
6						12	14	12	9					16.8
5					13	14	11	6*						16.5
4				12	14	11	9							16.7
3			13	15	13	9*								17.2
2		14	13	10										15.3
1	12	13	14							*N = 5-10				13.5
1-12	12.4	14.8	15.0	15.6	15.7	17.1	19.3	17.4	16.1	18.7	17.6	13.6	16.5	16.32

TABLE 6

MEAN MENTAL AGE (IN YEARS) BY AGE AND GRADE

(Not given where N is less than 5; marginal values based on all cases)

	Age													
	6	7	8	9	10	11	12	13	14	15	16	17	18	6-18
Grade														
12												17.0	17.0	17.2
11										18.5*	17.3	16.2	15.8	16.5
10									16.9*	16.6	15.3	15.4	14.8*	15.7
9								16.1*	15.1	14.8	14.7	15.2*		15.0
8							17.2*	14.9	14.1	13.0	12.7*			14.2
7							13.7	13.3	12.5	11.6*				13.2
6						12.8	12.3	12.2	11.2					12.4
5					11.2	11.2	11.1	10.3*						11.1
4				10.0	10.3	9.6	8.9							10.0
3			9.0	9.2	8.7	8.3*				*N = 5-10				9.0
2		7.6	8.0	7.8										7.8
1	6.2	6.8	7.1											6.5
1-12	6.2	7.2	8.2	9.4	10.4	11.5	12.4	13.5	13.8	14.7	15.5	16.0	16.2	

TABLE 7

STANDARD DEVIATIONS OF MENTAL AGES FOR AGE-GRADE DISTRIBUTIONS
(Not given where N is less than 5; marginal values based on all cases)

	Age													
	6	7	8	9	10	11	12	13	14	15	16	17	18	6-18
Grade														
12												1.4	2.4	2.1
11										1.8*	1.9	1.9	2.1	2.1
10									1.3*	1.8	1.6	2.2	2.0*	1.9
9								2.1*	2.0	2.1	2.2	1.7*		2.1
8							1.0*	2.0	1.7	2.0	1.9*			2.0
7							1.8	1.8	1.8	1.0*				1.8
6						1.3	1.7	1.6	1.3					1.6
5					1.3	1.5	1.3	.7*						1.4
4				1.1	1.4	1.2	1.1							1.3
3			1.0	1.3	1.3	1.0*					*N = 5-10			1.2
2		1.0	1.1	.9										1.1
1	.7	.9	1.1											.9
1-12	.7	1.0	1.2	1.4	1.6	1.9	2.3	2.3	2.2	2.7	2.6	2.0	2.5	

TABLE 8

MENTAL AGE RANGES BY SCHOOL GRADE

Grade	Mean M.A.	-1σ Lowest 16%	$+1\sigma$ Highest 16%	M.A. Range -1σ to $+1\sigma$	-2σ Lowest 2%	$+2\sigma$ Highest 2%	M.A. Range -2σ to $+2\sigma$
12	17.2	15.1	19.3	4.2	13.0	21.4	8.4
11	16.5	14.4	18.6	4.2	12.3	20.7	8.4
10	15.7	13.8	17.6	3.8	11.9	19.5	7.6
9	15.0	12.9	17.1	4.2	10.8	19.2	8.4
8	14.2	12.2	16.2	4.0	10.2	18.2	8.0
7	13.2	11.4	15.0	3.6	9.6	16.8	7.2
6	12.4	10.8	14.0	3.2	9.2	15.6	6.4
5	11.1	9.7	12.5	2.8	8.3	13.9	5.6
4	10.0	8.7	11.3	2.6	7.4	12.6	5.2
3	9.0	7.8	10.2	2.4	6.6	11.4	4.8
2	7.8	6.7	8.9	2.2	5.6	10.0	4.4
1	6.5	5.6	7.4	1.8	4.7	8.3	3.6

Chapter IV

URBAN-RURAL, OCCUPATIONAL, AND SIBLING RELATIONSHIPS

A quarter of a century ago such an accumulation of data as we can here present would have been hailed as definite proof that intellectual differences have an hereditary basis, but at the present time these data will not be regarded as of crucial significance in a field of controversy. The results, however, are of interest in that they provide rather definite information on certain concomitants of I.Q. variation, even though one cannot conclude therefrom that I.Q. variation is determined any more by hereditary factors than by the cultural milieu provided by parents.

All the data in this chapter except those for siblings are based upon penultimate scoring (see the first footnote in Chapter III). The age samplings, it will be recalled, were such as to yield a slight bias as regards occupational status of the father. We have no reason to suspect that within an occupational group any selective factors have been operative, except in the case of the Denver preschool group (see *Measuring Intelligence* , page 18), which is not being included in the occupational and urban-suburban-rural comparisons. The problem of securing samplings which will be representative of the urban, suburban, and rural populations is complicated by the possibility of differences between urban, or suburban, or rural communities. Our samplings are not adequate for the rural population — the extremely rural regions are not sufficiently well represented. It is because of this bias that Terman and Merrill were willing to tolerate I.Q. means in excess of 100 even after adjustment for the bias in sampling as regards occupational status (see *Measuring Intelligence* , Table 6).

To secure an adequate sampling of rural children is especially difficult because of the vast differences in so-called rural communities. For example, a certain California community has all the superficial characteristics of ruralism, but upon closer scrutiny it is found to be highly residential (retired people, long-distance commuters, ranch hobbyists, etc.) as compared to a rural region of the Dakotas or up-state New York or interior California. We stress the difficulty involved in rural sampling so that the reader will not draw any unqualified conclusions from the data on urban, suburban, and rural groupings.

I.Q.'s of Urban, Suburban, and Rural Children

The samplings for urban children and the number of cases by communities were as follows: Denver, 111 (excluding preschool as highly selected); Minneapolis, 183, New York, 48; Reno, 112; Richmond, Virginia, 187; San Antonio, 254; and San Francisco, 527. These cities should provide a fair urban cross-section. We are not reporting data separately by communities, but it can be noted that differences between the mean I.Q.'s for these cities tend to be small.

Into the suburban classifications we place the following communities: White Plains, New York, 160; Redwood City, California, 134; Los Gatos, California, 314; and four small communities just out of Kansas City in Johnson County, Kansas, with 199 cases drawn from Westwood View, Hickory Grove, Roseland, and Shawnee Mission schools. We admit to some arbitrariness in this grouping, but all these communities tend to be between urban and rural as regards their residential character and dependence upon larger near-by cities.

The samplings from rural communities include 85 from the Mount Washington School, Bullit County, and Liberty School, Oldham County, Kentucky. A total of 152 were drawn from the following districts of Indiana:

Prather School, Charlestown schools, and Borden High School in Clark County, Palmyra School and Morgan Township School in Harrison County, and Galena School in Floyd County. A farming region at Bloomington, Minnesota, supplied 92 cases; the farming and small village community of Randolph, Vermont, provided 275; and 65 subjects were secured in the vicinity of Atlee, Virginia. We have already expressed some skepticism concerning the representativeness of these communities.

Data on I.Q.'s for children in urban, suburban, and rural communities are presented in Table 9 for three age groupings. It will be seen that there is no appreciable difference between urban and suburban - a fact which should evoke no surprise. The mean for rural children of ages 2 to 5-1/2 is much nearer the urban and suburban averages than one would expect. As already implied, we believe that more adequate rural samplings would result in lower means than those given in Table 9.

TABLE 9

I.Q. DATA FOR URBAN, SUBURBAN, AND RURAL CHILDREN
(Denver 2 to 5-1/2 year-olds excluded)

	2 - 5-1/2			6 - 14			15 - 18		
	Urban	Suburban	Rural	Urban	Suburban	Rural	Urban	Suburban	Rural
N	354	158	144	864	537	422	204	112	103
M	106.3	105.0	100.6	105.8	104.5	95.4	107.9	106.9	95.7
σ	15.7	16.1	15.4	14.7	16.8	15.5	16.5	15.7	15.9

Occupational Differences

It has long been known that I.Q.'s of children tend to vary with the socio-economic status of their families, and that within any one socio-economic group there is considerable residual variation which is independent of socio-economic level. The material which can be presented here is extensive enough not only to confirm pre-

TABLE 10

L-M COMPOSITE I.Q.'S ACCORDING TO FATHER'S OCCUPATION

Father's Occupational Classification		2-5-1/2	6-9	10-14	15-18
I. Professional	N	36	31	41	16
	M	114.8	114.9	117.5	116.4
	σ	15.2	12.7	16.8	10.9
II. Semi-professional and managerial	N	50	52	75	38
	M	112.4	107.3	112.2	116.7
	σ	14.2	12.3	15.8	12.6
III. Clerical, skilled trades, and retail business	N	175	199	243	85
	M	108.0	104.9	107.4	109.6
	σ	14.6	14.7	16.4	15.5
IV. Rural owners	N	59	106	154	91
	M	97.8	94.6	92.4	94.3
	σ	15.0	13.7	15.9	17.8
V. Semi-skilled, minor clerical, minor business	N	224	249	289	104
	M	104.3	104.6	103.4	106.7
	σ	14.7	14.4	16.1	15.2
VI. Slightly skilled	N	59	77	90	32
	M	97.2	100.0	100.6	96.2
	σ	18.9	12.8	15.3	14.5
VII. Day laborers, urban and rural	N	30	58	67	27
	M	93.8	96.0	97.2	97.6
	σ	13.1	13.0	15.9	11.5

vious findings but also to permit a breakdown by age so that possible trends might come to light. We therefore present means and standard deviations (Table 10) for four different age groupings.

The Goodenough scheme of classifying on the basis of the occupational status of the father was followed. We hold no brief for this particular method of rating; it is as satisfactory or as unsatisfactory as one chooses to believe — we would prefer a better scheme. It should be noted, however, that such differences as are found between the average I.Q.'s for the seven classification

groups will have occurred in spite of the inadequacies of the classification scheme. The variability within a class is, of course, partly due to the heterogeneity of the occupations which go to make up the class.

The results as summarized in Table 10 need little discussion. The I.Q. means for children whose fathers are in the professional group tend to be about 18 or 20 points higher than the means for those whose fathers are in groups IV, VI, and VII. The overlapping of the highest with the lowest group is such that only about 10 per cent of the children in the latter group exceed the mean I.Q. for the former. There is also the fact that about 10 per cent of the children of professional men have I.Q.'s which are below the general average. This could be a genetically determined phenomenon which occurs despite the supposedly superior environment provided in the homes of the professional class. One also notes from Table 10 that the range of means is as great for the lowest age grouping[1] as for the later ages.

It is difficult to discern any trends with age in Table 10. Minor or chance fluctuations may be found where the N's are relatively small, and groups IV, VI, and VII tend to switch in relative positions. The low values for group IV would indicate that few, if any, bankers and large ranch owners have been included in the sample as rural owners.

Sibling Resemblances

The data for siblings which we can assemble are unique in several respects. (1) The I.Q.'s, being based on a composite of Forms L and M, are less subject to measurement errors than those used in previous studies

[1] These means will not check with those given in Table 12 of *Measuring Intelligence*, which were based on ages 2-1/2 to 5-1/2, erroneously transcribed as 2 - 5-1/2.

of sibling resemblances. (2) Scores on the same scale are available for sibling pairs with wide age differences for the two members of a pair. (3) By subdividing, it is possible to report correlations for preschool versus preschool sibling; for preschool versus older, in school, sibling; for young, say 6 to 11, versus older, 12 to 18, sibling; for pairs with both members between ages 6 to 11; ditto, ages 12 to 18, and for all sibling pairs regardless of age. Such correlations are in reality correlations between indices, but we show in Appendix A that there is no spurious element involved. (4) Our sibling samples may be regarded as being more representative than those used in previous studies.

It should be noted that twin pairs were not included as such, although each member of a pair was plotted against other siblings in the family. The correlations were computed from double-entry scatter plots except in those cases where selected younger were being correlated with older siblings. The standard errors for the resulting correlation coefficients have not been determined

TABLE 11

SIBLING RESEMBLANCES

	No. of families	No. of pairs	σ_x	σ_y	r
All possible	263	384	16.4	16.4	.53*
Ages 12 to 18 only	34	38	17.2	17.2	.54*
Ages 6 to 11 only	70	80	16.1	16.1	.57*
Ages 6 to 11 versus 12 to 18	80	104	18.1	15.6	.48
Ages 2 to 5-1/2 only	41	42	15.6	15.6	.55*
Ages 2 to 5-1/2 versus 6 to 18	81	119	15.9	15.1	.52

* Double-entry

for the simple reason that no formula is available which is adequate for the sibling situation involving a varying number of children per family. A maximal value will be yielded by using the number of families as N.

The results are summarized in Table 11. If these correlations were corrected for attenuation, they would run about .02 or .03 higher. Thus, without any further corrections, it can be said that our 384 pairs of siblings representing 263 families could be expected to show a resemblance of about .55 or .56 if measurement errors were not present. The more interesting facts, however, are the coefficients for preschool versus preschool (i.e. siblings of ages 2 to 5-1/2) and the preschool versus older siblings. Evidently the factors which tend to produce family resemblances in intelligence are not only operative at early ages but also continue to have an influence in maintaining that resemblance, or else new factors having the same effect are present.

Chapter V

SEX DIFFERENCES

It is not our purpose to discuss at length the importance of sex differences, nor shall we attempt an explanation of, or rationalization for, the interest of psychologists in the problem. One might, by some speculation, arrive at the notion that the reason for a section on this topic in so many research reports can be subsumed under one of the following three headings: a real interest in sex differences per se, or an interest incidental to the problem of uncontrolled variables in experimental work, or a mere following of the tradition of including a section devoted to sex differences. All three types of motivation have contributed to our knowledge concerning differences due to sex. Whether or not such differences as have been demonstrated have social significance may be open to question, but as regards their import for experimental control there can be no question; for when sex differences are not present, conclusions from an experiment become more general and less subject to qualifications.

One who would construct a test of intellectual capacity has two possible methods of handling the problem of sex differences. (1) He may assume that all the sex differences yielded by his test items are about equally indicative of sex differences in native ability. In this case separate norms for the sexes will be necessary if the means for total score on the battery show appreciable discrepancy for the sexes. (2) He may proceed on the hypothesis that large sex differences on items of the Binet type are likely to be factitious in the sense that they reflect sex differences in experience or training. To the extent that this assumption is valid he will be justified in eliminating from his battery the test items

which yield large sex differences, and by this method may be able to dispense with separate norms.

There are, of course, limits beyond which one would hesitate to go in defense of either of these methods. For example, the person inclined to favor separate norms for the sexes would nevertheless avoid using test items obviously unfair to one sex or to the other, such as making a dress or constructing a kite. On the other hand, one who favors the second procedure must admit that it rests upon an assumption, and that in the case of certain test items the assumption may be in error. Certainly the absence of sex differences cannot be proved by the simple expedient of refusing to use items which show such differences!

The authors of the New Revision have chosen the second of these alternatives and have sought to avoid using test items showing large sex differences in per cents passing. Their choice rests upon the empirical fact that test batteries of extensive scope and varied content as a rule yield only small sex differences in total scores, and that when individual test items do show large sex differences these can often be accounted for in terms of known differences in environment or training. However, because of the limited number of test items available at a given age level it was not possible to eliminate all of the items which showed sex differences. The extent to which sex-differentiating items were balanced will become evident in the ensuing section of this chapter. Subsequently, information will be given on those retained and eliminated items which yielded statistically significant sex differences.

Differences in I.Q. or Total Score

The success of the aim to produce a scale which will yield comparable I.Q.'s for the sexes may be seen by examining Table 12. The data in this table are based

TABLE 12

SEX DIFFERENCES IN I.Q.
(Composite of Forms L and M)

Age	N		Means		S.D.'s	
	M	F	M	F	M	F
2	54	47	105.5	108.0	17.1	12.6
2-1/2	49	53	105.7	115.5	17.3	20.8
3	50	49	107.9	104.6	16.6	20.1
3-1/2	52	51	106.6	114.0	17.3	14.1
4	50	55	105.1	105.4	13.5	17.9
4-1/2	51	50	103.7	106.3	15.2	15.4
5	52	57	103.2	107.2	14.5	12.7
5-1/2	57	52	101.6	101.2	13.4	14.3
6	102	101	102.0	100.5	13.2	11.6
7	102	100	103.4	101.2	15.3	16.1
8	102	101	104.3	101.8	14.4	16.2
9	103	101	106.1	103.2	16.2	15.8
10	101	100	104.2	103.8	15.4	16.3
11	102	102	104.4	104.1	17.2	17.9
12	102	100	103.4	102.6	20.7	18.6
13	103	101	104.9	102.0	16.8	18.5
14	98	102	101.4	100.0	16.6	15.9
15	51	56	106.0	99.3	19.9	17.9
16	51	51	101.0	101.2	17.1	17.4
17	54	55	106.6	103.6	14.3	13.1
18	50	51	106.4	108.3	17.7	15.5
2 - 5-1/2	415	414	104.8	107.8	15.8	16.9
6 - 14	915	908	103.8	102.1	16.4	16.5
15 - 18	206	213	105.2	103.0	17.4	16.4

upon L-M composite scores obtained by the final scoring procedures. The nearest approach to statistically significant differences between means will be found at ages 2-1/2 and 3-1/2 where the critical ratios are 2.6 and 2.4 respectively. These 10- and 8-point differences lose some significance because of a 3-point reversal at age 3. When ages 2 to 5-1/2 are combined (see bottom of table), the difference between the sex means is 3 points, which is 2.6 times its standard error. This suggests the likelihood that the scales are favorable to slightly higher scoring for girls at these early ages. For later ages the means for boys are rather consistently higher than those for girls. The average difference for ages 6 to 18 combined (1121 cases of each sex) is about 1.8, with a standard error of about .7, from which we cannot conclude that there is or is not a real sex difference in measured I.Q.'s at these age levels. We can be fairly confident, however, that the true difference is reasonably small, and consequently an obtained I.Q. need not be circumscribed because of sex. That intellect can be defined and measured in such a manner as to make either sex appear superior will become apparent in the next section.

Sex Differences by Items

One of the sources of conflicting data regarding sex differences in mental ability must be attributed to differences in tests. Certainly, tests which bear the same label are apt to be quite different as to content and as to the kind of ability called for. One does not have to pursue this line of thought very far to realize that all the issues revolving around the organization of abilities are here involved. The problem would, perhaps, be greatly clarified if the factor analysts should succeed in two things: first, the isolating and unequivocal tagging of abilities; and second, the constructing of tests by which such 'pure' traits or abilities can be measured independently.

With our present state of knowledge (mostly ignorance) concerning just what abilities exist and which test measures what, it may be more profitable to study sex differences via the individual items. Such a procedure will overcome one of the chief limitations to the study of sex differences by way of scores based upon a composite of items, namely the likelihood that such differences as emerge may be due to a particular cluster of items which call for some factor or ability unknown to or unrecognized by the investigator. There is also the possibility that a composite (aggregation of items) may be so balanced, either by design or accident, as to mask real differences. Now, both an item and a composite will likely call for a complex of abilities, but the factorial composition of a single item should be simpler than that of a composite unless the latter has been constructed as a 'pure' measure of some one ability or unless it consists of an aggregation of highly similar items such as, for example, the verbal analogies subtest of the typical group scale of intelligence. A knowledge of just what types of item situations reflect sex differences should also make it easier to theorize as to possible causes of the differences.

But these advantages of the item approach are somewhat offset by the resultant large mass of data which may be too unwieldy to permit of generalizations. Furthermore, such generalizing as one attempts will likely involve some sort of reference frame as regards underlying abilities. The reference frame adopted may be one which has been arrived at empirically, say by factor analysis, or it may be one's own arbitrary, rational, albeit subjective, ideas regarding abilities. In either case, generalizing may involve something akin to word magic, but if the original data are reported completely, any reader can readily use whatever frame of reference he chooses: he can become entranced by his own word magic.

The analysis of item data for sex differences must depend upon the per cent passing and failing each item.

SEX DIFFERENCES

The revision data are such that a comparison for a single item can be made for several different age groups; thus for each item one could compute several critical ratios for the sex difference in per cent passing. It would seem desirable to have a single index for the statistical reliability of the differences on an item. Obviously, one cannot justifiably combine ages so as to have a single per cent for each sex, nor can one average the several age per cents. A suitable single index can be obtained by computing, for each age, chi square from the fourfold table formed by classifying passing and failing by sex, then summing the chi squares for the several ages. The number of degrees of freedom will equal the number of chi squares summed, i.e. the number of ages for which separate values of chi square are determined. The number of age levels usable for a given item will depend upon the spread of failing and passing for the item, but the extremes cannot be used because of the inapplicability of chi square when any one expected frequency in the fourfold table is small.

TABLE 13

EXAMPLE OF CHI SQUARE AS APPLIED TO TESTING THE SIGNIFICANCE OF SEX DIFFERENCES

Age	6			7			8			9		
	+	-										
B	18	84	102	36	66	102	44	58	102	66	37	103
G	8	93	101	20	80	100	39	62	101	49	52	101
	26	177	203	56	146	202	83	120	203	115	89	204
x^2	4.30			5.89			.43			5.02		

Sum x^2 = 15.64; n = 4; P = .0036

SEX DIFFERENCES

As an example of the use of the chi square technique, let us look at Table 13, wherein will be found the frequency of passing (+) and failing (-) for each sex by separate ages for the item 'orientation: form.' Under each fourfold table will be found the chi square for that table. Taken singly no one chi square is significant (it will be recalled that a chi square based on one degree of freedom corresponds to the square of a critical ratio), but all four differences are in the same direction, so one need not be surprised at the significant x^2 probability of .0036, which the reader will recognize as being near the P yielded by a critical ratio of 3.

The use of chi square does not eliminate the question as to what shall be accepted as a critical value for significance versus insignificance. A P of .05 may be sufficiently low to suggest non-chance differences, but a smaller value should be demanded before one can be very sure of the reality of the obtained differences. We shall report here only on those items which yield P values of .01 or less. Smaller values of P will tend to give us still greater confidence for concluding that a real sex difference exists. The chi square technique, like the ordinary critical ratio method, does not yield a measure of the degree of association, and consequently it cannot be inferred from two chi square P values that the association is necessarily stronger for the item yielding the smaller P.

The essential data on item sex differences are summarized in Tables 14 and 15. These tables include items which were eliminated and items which were retained. The locations of the latter in the final forms are indicated. The items have been arranged according to the age groups utilized in the sex comparisons, and these ages have been specified in the tables. The reader will recall that half-age groups, 2-1/2, 3-1/2, 4-1/2, and 5-1/2, were tested, and that there are approximately 50 of each sex in age groups 2 to 5-1/2 and 15 to 18, and about 100 of each sex in age groups 6 to 14. More exact N's for the data on any one item may be obtained by

combining the information in Tables 14 and 15 regarding ages used and the several age-sex N's given in Table 12.

A word is in order regarding the varying ranges of ages used. The requirement that expected frequencies, for passing or failing, be 10 or greater constitutes the principal restriction for usable ages. For a few items, as for example the sample item in Table 13, an additional restriction was involved which had to do with the faulty location of the items in the tryout forms. These restrictions will not tend to enhance or diminish sex differences, but will heighten our confidence in the adequacy of the data used in the comparisons. The two 'plan of search' items constitute the only exceptions to the foregoing procedure. For both these items, there were very minor sex differences prior to age 13, then marked and consistent differences from 13 to 18; hence we report data only for these later ages, and consequently any conclusion must be modified accordingly. Incidentally, these two items and two of the 'copying a bead chain from memory' items were the only items of all those in the provisional tryout forms which seemed to suggest the emergence of a sex difference with age. In all other cases the differences were either consistent and significant or inconsistent and insignificant as regards the comparisons at the several ages.

It should also be noted that data for a recurring test are presented for the test only at one passing level unless it is possible, as in the case of one test — 'orientation: form' — to make comparisons for a different set of age groups. For instance, test L, XIV, 4, 'ingenuity,' which is scored as 1 plus at this level and for which sex comparisons can be made for ages 11 to 17, recurs as L, A.A., 6 with score as 2 plus, but the possible sex comparisons would involve ages 12 to 18. Obviously, the presentation of such duplicative data is not only unnecessary but actually questionable. As a matter of record, it can be stated that when a recurring test showed a significant sex difference for one passing standard, the dif-

TABLE 14

ITEMS ON WHICH GIRLS SURPASS BOYS

Location	Name	Ages	x^2, P
	Picture memories	2 -3	.0046
	Counting (reciting numerals)	$2\frac{1}{2}$-4	.0007
	Paper folding: square	$2\frac{1}{2}$-4	.00024
	Buttoning	3 -4	.0013
L, IV-6, 1	Aesthetic comparison	3 -5	.00018
	Aesthetic comparison	$3\frac{1}{2}$-$5\frac{1}{2}$	.0014
M, III-6,a	Matching objects	$3\frac{1}{2}$-5	.010
L, V, 2	Paper folding: triangle	4 -5	.0010
	Tying a bow knot	$5\frac{1}{2}$-8	.000014
	Age discrimination	8 -12	.0006
M, XI, 2	Copying a bead chain from memory	7 -17	.0065
L, XII, 6	Minkus completion	10 -17	.0010
M, A.A., 4	Codes I	10 -18	.0065

ference was also present for all passing standards.

We are now ready to discuss such sex differences as exist for single items. First we note from Tables 14 and 15 that there are apparently more items on which boys surpass girls than vice versa, except at the pre-school levels. This will account for part of the small I.Q. superiority of girls at the lower ages and of boys at the middle and upper ages. But the greater number of items favoring boys is more apparent than real. Actually there are only 14 item situations involving male superiority as compared to 16 item situations for which girls are superior. This repetition of item situations, i.e. variant forms for certain items, by which 25 items are reduced to 14 item situations is important as regards the study of sex differences per se, but as regards the New Re-

vision as a measuring instrument the inclusion of variant forms of items which yield such sex differences may be regarded as unfortunate insofar as their presence has not been entirely balanced by items favorable to the opposite sex (see Table 12).

The interpretation of the findings summarized in Tables 14 and 15 is fraught with difficulties. Most of the superiority for girls occurs at the lower age levels on items which seem to involve some type of manipulative ability ('buttoning,' 'tying knot') or discrimination ('aesthetic comparison,' 'picture memories') or number facility ('counting,' 'matching objects'). The direction of the difference for 'buttoning,' 'tying knot,' and 'aesthetic comparison' ('Which one is prettier?') is plausible, but the differences on the other items at these lower ages are baffling.

The superiority of girls on 'age discrimination' (from pictures) fits in with the notion that girls are more interested than boys in people and social matters. The difference on 'copying a bead chain from memory' is in reality small - a comparison on about 950 of each sex is necessary to produce the not so significant P of .0065. The superiority of the girls on this item is somewhat greater and more significant for ages 11 to 17 than for 7 to 17; there is some indication of emergence at age 11, a trend also present for the parallel item in Form L (XIII, 6) which yields non-significant differences in the same direction as M, XI, 2. The similar and easier item at L, VI, 2 shows no sex difference; neither does the non-memory 'copying of a bead chain' (M, VI, 2). The meaning of the superiority of the girls on 'Minkus completion' is open to question since the parallel item on Form M shows only a slight difference for the same age samplings. Likewise, one wonders what psychological significance can be attached to the difference on 'codes' since the data for the other three code items, two retained and one not retained in the final forms, show an utter lack of sex difference.

We have remarked earlier that a part of the inconsistencies among the data on sex differences is due to tests with the same label possibly being measures of different abilities. We now see that highly similar items do not yield consistent sex differences. We are unable to see any reason why a difference should emerge on this particular 'Minkus completion,' or 'codes,' or 'bead chain' item, and not on parallel items.

Let us now turn to the items upon which boys tend to excel (Table 15). The most striking fact shown in this table is the marked difference, highly significant and consistent with age, on 'picture absurdities.' That absurdity per se is not the sex-differentiating factor is evident from the absence of a difference for 'verbal absurdities.' The results for tests involving 'orientation' are also striking and perhaps not so surprising. It should be reported that 'orientation: direction II' on Form M, with frequencies too low for adequate statistical treatment, exhibits the same trend. It may be that boys are superior in space ability, and that the 'orientation' items, along with 'block counting' (from pictures) and 'plan of search,' depend somewhat on such an ability. The 'substitution' item as here used might also involve space. It might have been anticipated that the 'induction,' 'arithmetical reasoning,' and 'ingenuity' items, and certainly 'word naming: vehicles,' would favor the males.

The three items of Table 15 not yet discussed are of especial interest in that parallel, or similar items, fail to show similar differences. Of the 7 'opposite analogies' items, only one shows a sex difference. It contains the sub-items; 'the rabbit's ears are long, the rat's are_____; snow is white, coal is_____; the dog has hair, the bird has_____; wolves are wild, dogs are_____.' Of 7 'comprehension' items, only one yields a significant difference. One of its three questions is 'What makes a sailboat move?' Of 4 sets of 'abstract words,' the one upon which boys are superior requires definition for the words 'connection,' 'compare,'

TABLE 15

ITEMS ON WHICH BOYS SURPASS GIRLS

Location	Name	Ages	x^2, P
	Picture absurdities	4 -6	.0026
L, VII, 1	Picture absurdities I	$5\frac{1}{2}$-9	<.000001
M, VII, 3	Picture absurdities I	6 -9	.0010
	Orientation: form I (1+)	6 -9	.0036
L, VIII, 5	Comprehension IV	7 -10	.00026
	Opposite analogies	7 -12	.0010
M, X, 1	Block counting	7 -12	.0007
	Orientation: form II	9 -14	.00011
L, X, 2	Picture absurdities II	9 -14	<.000001
M, XII, 5	Picture absurdities II	9 -15	.00002
L, XIV, 3	Picture absurdities III	9 -17	<.000001
	Word naming: vehicles	9 -17	<.000001
M, XIV, 3	Orientation: direction I	9 -17	.00005
M, XIII, 2	Memory for stories II: acrobat	9 -18	.0002
	Orientation: form I (2+)	9 -18	.00001
L, XI, 3	Abstract words I	10 -15	.0070
L, XIV, 5	Orientation: direction I	10 -17	.000003
L, XIV, 4	Ingenuity	11 -17	.000001
L, XIV, 2	Induction	11 -18	.000003
L, A.A., 4	Arithmetical reasoning	11 -18	.0092
M, XIV, 5	Ingenuity	12 -17	.0016
	Substitution	12 -18	.00008
L, XIII, 1	Plan of search	13 -18	.0082
M, XIII, 1	Plan of search	13 -18	.0010
L, S.A. III, 2	Orientation: direction II	15 -18	.0040

'conquer,' 'obedience,' and 'revenge.' It happens that a parallel item containing the words 'pity,' 'curiosity,' 'grief,' and 'surprise' shows slight, though consistent from age to age, differences in favor of girls.

The results for these three types of items ('comprehension,' 'abstract words,' and 'opposite analogies') plus similar inconsistencies for some items showing female superiority, tend to point to one conclusion: namely, that sex differences are apt to be a function of the content of an item rather than of any basic abilities called for by the item. In other words, observed sex differences in scores may be either highly specific to an item situation or a reflection of a real difference in ability.

Chapter VI

DATA ON RELIABILITY

It is the purpose of this chapter to set forth more information on the reliability of the scales than was feasible in the previous volume. As stated therein, the ordinary form versus form coefficient of reliability is inapplicable to I.Q. data because of an apparent lack of homoscedasticity in the scatter plots. This tendency for the scatters to be fan-shaped was evident in nearly all the scatter diagrams, so that it was assumed to be a real, non-chance phenomenon. This was taken to indicate that the reliability of an I.Q. score is a function of the magnitude of the I.Q. itself, and in order to determine the error of measurement associated with a given I.Q., resort was made to estimation by way of the average difference between I.Q.'s derived from Form L and Form M. The resulting reliability data reported in the previous volume were confined to large I.Q. groupings for ages 3 to 18 combined.

More specifically, we propose in this chapter (1) to present evidence to show that the lack of homoscedasticity is non-chance; (2) to break down the data into three age combinations with smaller I.Q. groupings; (3) to supplement the average-difference method, which assumes normality of distribution for the errors of measurement, by another scheme which is not subject to this limitation; (4) to state more fully why we believe that the observed facts regarding the dependence of the errors of measurement upon the magnitude of the I.Q. might have been anticipated on logical grounds; and (5) to provide information on the accuracy of mental age scores.

When the several reliability scatter plots were made for the 21 age groups, a tendency toward a fan type of distribution was noticeable in 17 of the plots. In

order to get a statistical measure of this heteroscedasticity, the differences between Form L and Form M I.Q.'s were plotted against the composite I.Q.'s, obtained by combining results from the two forms. The computed coefficients of correlation between I.Q. differences and I.Q. magnitude were positive in 18 instances, and ranged from -.103 to .240 with a median value of about .16. Only one of the separate values could be deemed significantly different from zero, but the consistently positive coefficients from independent samples suggest that something more than chance is operating here. When ages 2-1/2 to 18 are combined, the correlation is .135, with a standard error (by the orthodox formula) of .018. This coefficient is 7-1/2 times its sampling error and therefore definitely greater than zero. In fact, we can be reasonably sure that the universe value for r is greater than .08.

TABLE 16

REGRESSION OF I.Q. DIFFERENCES, y, ON COMPOSITE I.Q., x

Age	N	M_x	σ_x	M_y	σ_y	r_{xy}	b_{yx}
2-1/2 - 5-1/2	728	106.27	15.93	6.00	4.78	.108	.145
6 - 13	1623	103.69	16.60	5.03	3.92	.131	.191
14 - 18	619	102.97	16.77	4.47	3.50	.148	.205
2-1/2 - 18	2970	104.17	16.52	5.15	4.10	.135	.156

In Table 16 will be found pertinent data on the relationship between I.Q. differences and I.Q. magnitude for three age combinations and for the total group. (Age group 2 has not been included because the I.Q.'s of two-year-old children cannot fall below 75, a factor which tends to limit the difference between Form L and Form M I.Q.'s for those who are retarded.) Despite the smaller N's for the three age groups, the three correlations are too large to be considered chance deviations

from zero. These facts tend to substantiate the conclusion that the lack of homoscedasticity in the original reliability plots is statistically significant. That such small correlations as .135 can possess more than statistical significance will become apparent and real in the pages to follow.

An examination of the four scatter diagrams from which the data of Table 16 were derived indicated that the regression of I.Q. differences on I.Q. magnitude might not be linear; so etas for differences on I.Q. were computed. In order to test for lack of linearity, Fisher's[1] analysis of variance method was used. For the 2-1/2 to 5-1/2 group there is no evidence of real curvilinearity; for ages 6 to 13, Fisher's test indicates that the probability of as great a departure from linearity is less than .001, and accordingly an assumption of linearity is not tenable; for ages 14 to 18, the observed discrepancy from linearity would arise by chance 5 times out of 100, which suggests a real departure from linearity. For ages 2-1/2 to 18 combined the probability of as great a deviation from linearity is .01. In view of these probability figures it does not seem safe to assume that the relationship between I.Q. size and I.Q. differences, i.e. I.Q. accuracy, is linear. The curvilinear trends will be described subsequently and factors which may possibly explain the lack of linearity will be suggested.

It is convenient at this time to restate why we think it very logical that the standard error of measurement for I.Q.'s should vary in such a way that greater accuracy is associated with lower than with higher I.Q.'s. Suppose we have an individual whose M.A. is 100 months and we accept 4.42 months as the standard error of measurement to be associated with a mental age of 100 months. If the individual's C.A. were 100 months, his I.Q. would be 100 $\pm$ 4.42; if his C.A. were 80, his I.Q.

[1] R. A. Fisher, *Statistical Methods for Research Workers* (6th ed.), pp. 257-261. Edinburgh: Oliver and Boyd, 1936.

would be 125 ± 5.52; and if his C.A. were 125, his I.Q. would be 75 ± 3.54. (These three standard errors of measurement for the I.Q.'s follow from the principle that if a measure is transformed by division, its error must also be likewise transformed.) Now suppose that three individuals have mental ages of 96, 120, and 144 months. It seems reasonable to associate with each of these three mental ages that error of measurement which is found for individuals who usually score at these three levels. The proper errors would therefore be the errors found for individuals of C.A. 96, 120, and 144 respectively, and even though the 'reliability' coefficients were the same for these three age groups, the errors of measurement for M.A.'s would be different because of the increasing variability of the M.A. distributions as we pass from C.A. group 96 to 120 to 144. The respective observed errors in months are approximately 4.4, 5.2, and 5.9. If the C.A.'s for all three individuals were 120, their I.Q.'s would be 70 ± 3.7, 100 ± 4.3, and 120 ± 4.9. It might be argued that it is proper to consider only 5.2 as the error of M.A. determination for individuals of C.A. 120, but this would lead us astray since the reliability scatter plot of mental ages for a single chronological age shows the same fan-shape as the I.Q. scatter plot.

Perhaps an analogy will help clarify the issue. Suppose we were measuring the heights of individuals with a one-inch ruler, and that it is shown empirically that the shorter are measured more accurately than the taller children. No one could object to attaching to a child's height, regardless of his age, that error which is usually associated with like heights. If the given height in inches were transformed to some index of height, the error in inches must also be transformed into index units. Now, the increase in scatter of M.A. distributions with age obviously means that the error of measurement for M.A.'s also increases, and therefore the higher the M.A. score, the larger its absolute error. Our argument is predicated upon the idea that the error

of measurement for a given M.A. is not entirely a function of the age of the child but is a function of his level of performance in terms of mental age.

Let us now consider the two methods used to obtain the reliability of I.Q. scores. In view of an appreciable difference in reliability as we pass from the preschool to the higher age levels, we have broken down the total into three age combinations: 2-1/2 to 5-1/2, 6 to 13, and 14 to 18. The grouping together of several ages is necessary, as will soon become apparent, in order to have a sufficient number of cases for determining adequately the accuracy of I.Q.'s between, for example, 70 and 80; but the particular age combinations which we have used are admittedly arbitrary.

The first method of ascertaining reliability involved computing the average difference between Form L and Form M. I.Q.'s for those individuals whose composite I.Q.'s fall in a given interval, e.g. 70 - 79, and then multiplying this mean difference by $\frac{1}{2}\sqrt{\pi}$ to obtain the standard error of measurement, σ_e, from which the equivalent reliability coefficient is obtained via the relationship $r_{11} = 1 - \sigma_e^2/\sigma^2$ where σ^2 is the variance for the distribution of I.Q.'s. This method is based on the assumption that the errors of measurement are normally distributed, and that the two observed I.Q.'s for each of several individuals having the same 'true' I.Q. can be considered as though drawn by chance from such a normal distribution with unknown, but to be determined, variance, σ_e^2. It has been shown [1] that when the standard deviation for a variable is known and pairs of scores are drawn at random, the expected average difference between pairs will be $(2 \div \sqrt{\pi})$ times σ. In the present problem, we have an observed average difference from which to estimate σ. That this method of estimating a reliability coefficient is really satisfactory has been

[1] Q. McNemar, "The Expected Average Difference Between Individuals Paired at Random," *J. Genet. Psychol.*, 1933, 43, 438-439.

checked empirically by computing the ordinary form versus form reliability coefficients from 24 scatter plots (varying N's), and then recomputing the coefficients via the 24 mean differences between pairs of scores in a set (no grouping on the basis of composite scores involved). The average discrepancy between the coefficients obtained by the two methods was .005, the maximum discrepancy was .01, and there was no evidence for a directional bias. We conclude, therefore, that the average difference method is defensible even though based upon an assumption which may not be tenable.

In applying this method, those individuals whose composite I.Q.'s were 69 or less were grouped together, those above 140 were treated as a group, and the rest were grouped into intervals of 10 I.Q. points as 70 - 79, 80 - 89, etc. The average difference between L and M I.Q.'s was computed for each I.Q. group, and from these averages the *equivalent* reliability coefficients, as given in the left-hand column of Tables 17, 18, and 19, were estimated. A standard deviation of 16.6 has been used in all cases, and therefore these 'reliabilities' are comparable as far as range is concerned. This σ of 16.6 is the average of the two sigmas for the L and M I.Q. distributions for 2970 cases, ages 2-1/2 to 18. Had the value of 16.4 (used in the previous report) been employed again, the herein given coefficients would be on the average about .002 smaller. Further discussion of the results in Tables 17-19 will be postponed until the other method of determining the reliabilities has been described.

This second method depends essentially on determining, by direct computation of array variances, how accurately I.Q.'s on one form can be predicted from I.Q.'s on the other form, and then asking what the equivalent correlation (in this case, reliability coefficient) must be for such an error of estimate or array variance. Scatter plots were made with an interval of size 2 (small so as to avoid grouping error) for Form L I.Q.'s as ordinate and an interval of size 5 for M I.Q.'s as abscissa.

From these plots (again age groups must be combined so as not to have too few cases per array) the variances for the several vertical arrays about their respective means were computed. This yielded a separate variance for each of the 60 - 64, 65 - 69, etc. arrays. Scatter plots were also made with the axes reversed so that Form M I.Q.'s were along the y-axis, intervals of size 2, with the L values along the x-axis, 5-point intervals. From these plots a second set of array variances was obtained by the procedure just described. Thus were provided two variances for the 60 - 64 array, two for the 65 - 69 array, and so on, but the two variances differed by chance, as did also the number of cases in the two corresponding arrays. It would seem reasonable to average the two variances as a better estimate for a given array.

The procedure followed, however, was the averaging of four variances for, say, the 60 - 64 and 65 - 69 intervals or arrays so as to obtain an estimate of the variance for the 60 - 69 array. This final variance is a weighted average and is therefore equivalent to combining sums of squares of deviations from the four respective array means and then dividing by the total number of cases in the four arrays. This apparently roundabout procedure for obtaining the variances for the 60 - 69, 70 - 79, etc. arrays was employed in preference to the usual method (i.e. combining array distributions) in order to reduce somewhat the effect of regression within an array. This tendency for regression to increase the array variances has not, however, been completely eliminated; the computed, and therefore average, variances will be slightly exaggerated, and accordingly the estimated equivalent reliability coefficients will possess a small negative bias.

As has already been indicated, each of the several array variances will correspond closely to the square of the standard error of estimate in predicting I.Q.'s on one form from I.Q.'s on the other. The fact that these variances differ systematically as we pass from array to

array was expected because of the apparent heteroscedasticity of the scatter plots. The estimated equivalent reliability coefficient is easily determined from the relationship $r^2 = 1 - \sigma_a^2 / \sigma^2$ where σ_a^2 equals the given array variance and σ is 16.6. The estimates obtained by this method will be found in Tables 17-19, and the respective N's are given in two columns since the intervals, 60 - 69, etc. on the Form M axis will ordinarily not contain the same number of cases as the corresponding intervals on the Form L axis. It should also be noted that the total N's in these two columns need not agree mutually or with the total N for the average-difference method. This is due to the fact that the terminal intervals for the variance method were 60 - 69, and 140 - 149, whereas for the difference method these were 69 down and 140 up. The difference method utilizes all the cases, whereas in the variance method no attempt was made to use arrays below 60 or above 150 because of too few cases. These end intervals might have been handled otherwise, with a resulting negligible change in final values.

TABLE 17

RELIABILITIES FOR AGES 2-1/2 to 5-1/2

	Via av. diffs.		Via variances			Av.	Smoothed	
I.Q.	r_{11}	N	r_{11}	N_L	N_M	r_{11}	r_{11}	σ_e
140 - 149	.850	15	.796	11	11	.823	.834	6.8
130 - 139	.801	29	.912	41	33	.856	.849	6.5
120 - 129	.870	100	.866	82	96	.868	.874	5.9
110 - 119	.909	160	.888	175	168	.898	.890	5.5
100 - 109	.900	198	.908	191	184	.904	.899	5.3
90 - 99	.898	125	.892	108	126	.895	.909	5.0
80 - 89	.930	61	.924	76	63	.927	.914	4.9
70 - 79	.926	29	.916	20	28	.921	.919	4.7
60 - 69	.909	11	.911	7	4	.910	.914	4.9

DATA ON RELIABILITY

TABLE 18

RELIABILITIES FOR AGES 6 to 13

I.Q.	Via av. diffs. r_{1I}	N	Via variances r_{1I}	N_L	N_M	Av. r_{1I}	Smoothed r_{1I}	σ_e
140 - 149	.944	34	.903	29	32	.924	.907	5.1
130 - 139	.851	69	.896	73	57	.874	.898	5.3
120 - 129	.904	163	.890	139	171	.897	.897	5.3
110 - 119	.921	313	.921	309	316	.921	.916	4.8
100 - 109	.939	385	.922	407	400	.930	.926	4.5
90 - 99	.925	339	.930	333	341	.928	.932	4.4
80 - 89	.933	219	.943	216	196	.938	.941	4.0
70 - 79	.962	68	.950	76	68	.956	.958	3.4
60 - 69	.986	33	.972	30	28	.979	.971	2.8

TABLE 19

RELIABILITIES FOR AGES 14 to 18

I.Q.	Via av. diffs. r_{1I}	N	Via variances r_{1I}	N_L	N_M	Av. r_{1I}	Smoothed r_{1I}	σ_e
140 - 149	.942	8	.964	6	11	.953	.948	3.8
130 - 139	.950	24	.926	26	30	.938	.939	4.1
120 - 129	.930	73	.920	79	63	.925	.930	4.4
110 - 119	.928	118	.928	121	126	.928	.925	4.5
100 - 109	.925	125	.920	110	134	.922	.933	4.3
90 - 99	.948	142	.950	149	124	.949	.944	3.9
80 - 89	.966	80	.957	77	79	.961	.957	3.4
70 - 79	.959	35	.960	39	34	.960	.970	2.9
60 - 69	.992	14	.984	9	16	.988	.979	2.4

A comparison of the equivalent reliability coefficients (see Tables 17-19) reveals a median discrepancy of .01 between the values obtained by the two methods. We have no way of evaluating these differences statistically, but their magnitudes are such that it does not seem unreasonable to attribute the discrepancies to chance. There is one disturbing disagreement in Table 17, for the 130 - 139 level, for which, if non-chance, we have no explanation. It will also be noted that the intra-pair differences in each table are small relative to the range of coefficients; i.e. the two estimates tend to vary together. It would seem that a combination of the values obtained by the two methods would yield a more dependable estimate, hence the column in each table headed Av. r_{11}. These last values have been smoothed by the method of moving averages (over 3 observations) on the defensible assumption that the irregularities are partly due to chance. The smoothed values, and the corresponding standard errors of measurement, are given in the last two columns of Tables 17-19.

We consider these final values to be reasonably accurate estimates of the equivalent reliabilities and of the errors of measurement, to be sufficiently refined for all practical purposes, and to be decidedly superior to any computed reliabilities which ignore the lack of homoscedasticity in the scatter plots. Personal preference and utility must be the deciding factors as to whether a reliability coefficient or the corresponding error of measurement is considered of more fundamental importance. The latter is a measure of absolute error; the former of relative accuracy. Our final reliability coefficients are to be interpreted exactly as usual except that they refer to the accuracy of scores for a given I.Q. level relative to the total spread of I.Q.'s. They are not to be erroneously interpreted as giving the accuracy relative to a narrow range of I.Q.'s such as 80 - 89.

Plots of the smoothed reliability coefficients against I.Q. magnitude, one for each of the three age

groups, reveal apparently curvilinear relationships. The same holds true for the standard errors of measurement versus I.Q. level, as can be seen in Figure 3. On the basis of our argument that the errors of measurement for I.Q.'s should vary with I.Q. magnitudes because of increasing variability in M.A. distributions with age, we should anticipate a linear relationship between the σ_e's and I.Q. magnitude. This assumes that the reliability of the I.Q. scale does not vary with age and that the spread (standard deviation) of M.A. distributions increases with age at a constant rate (.166 times C.A.; i.e. this assumption implies that our computed value, namely 16.6, for the σ of the I.Q. distribution holds true for all ages). The three straight lines in the figure indicate the presumed relationships.

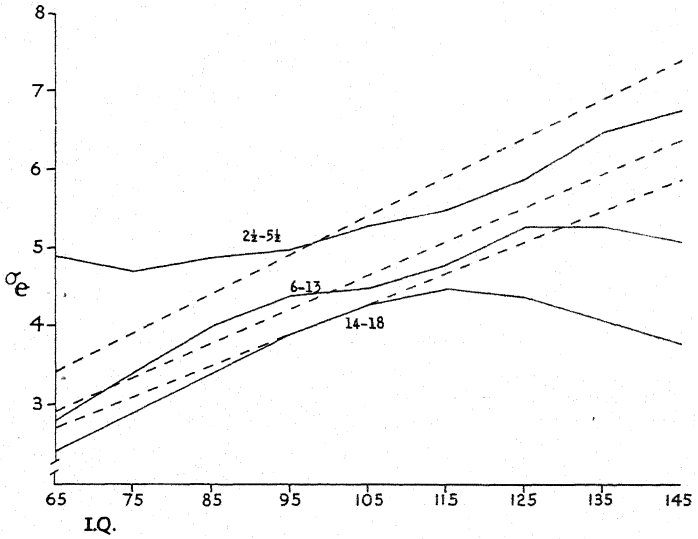


Fig. 3. Observed and expected (dotted) values for σ_e (based on Table 20)

DATA ON RELIABILITY

Let us now compare our final errors of measurement for I.Q.'s with those that one would expect to arise solely on the basis of the increase with age of the variability of M.A. distributions. If, for example, we choose the 6 to 13 group and if we accept .929 as the equivalent reliability coefficient and 4.42 as the corresponding standard error of measurement for I.Q.'s of 100, we can obtain the expected standard error for any given I.Q. by multiplying it by 4.42. Thus for an I.Q. of 125 we should expect the error to be 5.5 as compared with an observed value of 5.3; the expected value for an I.Q. of 75 would be 3.3, whereas the observed value is 3.4. These and the other values for ages 6 to 13 have been set forth in Table 20, in which will also be found corresponding figures for the two other age groups. The expected values for the lowest age group are based on a σ_e for I.Q. = 100 of 5.14, while for the 14 to 18 group, 4.10 was used.

TABLE 20

COMPARISON OF OBSERVED (O) WITH EXPECTED (E) ERRORS OF MEASUREMENT

	$2\frac{1}{2}$ - $5\frac{1}{2}$		6 - 13		14 - 18	
I.Q.	O	E	O	E	O	E
140 - 149	6.8	7.4	5.1	6.4	3.8	5.9
130 - 139	6.5	6.9	5.3	6.0	4.1	5.5
120 - 129	5.9	6.4	5.3	5.5	4.4	5.1
110 - 119	5.5	5.9	4.8	5.1	4.5	4.7
100 - 109	5.3	5.4	4.5	4.6	4.3	4.3
90 - 99	5.0	4.9	4.4	4.2	3.9	3.9
80 - 89	4.9	4.4	4.0	3.8	3.4	3.5
70 - 79	4.7	3.9	3.4	3.3	2.9	3.1
60 - 69	4.9	3.3	2.8	2.9	2.4	2.7

It will be noted, from the figure and Table 20, that there is a fair agreement between the linear expected and the observed values for ages 6 to 13, the group which contains the largest number of cases and for which the assumptions underlying the determination of the expected values are most nearly met. The failure, if non-chance, of the observed values for the 130 - 139 and 140 - 149 levels to agree with the expected values may be due to the brighter 11-, 12-, and 13-year-olds having mental ages in those levels where there is no longer an increase in M.A. variability. When the two sets of values for the 14 to 18 group are compared, we see that the agreement is close for I.Q. levels below 110, whereas for higher levels the accuracy does not decrease as expected. The explanation for this is that there is no increase in M.A. variability from ages 15 to 18, and therefore the factor which would, according to our reasoning, lead to a decrease is not here present. The agreement for ages 2-1/2 to 5-1/2 is not particularly striking, but two factors enter here which not only prevent concurrence but also act in such a way as to provide explanations for the direction of the discrepancies. One of these factors is the apparent, and we fear real, decrease in I.Q. spread as we pass from ages 2-1/2 to 5-1/2. This means that the variability of the M.A. distributions is not increasing as rapidly as assumed in determining the expected values for σ_e; hence the change in the observed σ_e's with I.Q. as we proceed upward and downward from the 100 I.Q. level will be less than anticipated. (As a matter of fact, a revision of our expected values to allow for a smaller constant increase in M.A. spread definitely swings, i.e. reduces the slope of, the line of expected values so as to reduce the discrepancies). The other disturber operating in the 2-1/2 to 5-1/2 group is the fact that the scales are progressively more reliable (as judged by form versus form coefficients, adjusted for differences in I.Q. range) as we pass from age 2-1/2 to 5. This also operates so as to yield errors of measure-

ment smaller than expected for higher I.Q.'s and larger than expected for lower I.Q.'s.

In view of the general marked agreement between expected and observed accuracy, and the very plausible explanations for such discrepancies as occur, we are inclined to repeat with still greater confidence the statement made in the previous volume to the effect that the dependence of I.Q. accuracy upon I.Q. magnitude is inherent in the I.Q. technique and should have been anticipated on strictly logical grounds.

TABLE 21

AVERAGE DEVIATIONS FOR TEST-RETEST OTIS I.Q.'S AS FOUND BY HIRSCH

I.Q. levels	N	Av. dev.
131 up	29	7.1
116 - 130	55	5.5
95 - 115	174	5.3
94 down	85	4.3

It is extremely likely that this factor accounts for the larger fluctuations of I.Q.'s, in test-retest or constancy studies of the 1916 Revision, for superior as compared to average or below-average individuals. That the same factor may be operative in group tests is suggested by the follow-up study of Hirsch,[1] who used the Otis Primary and Otis Advanced Tests. He reports the average deviation in test-retest I.Q.'s for four I.Q. groups, and his results, given in Table 21, are certainly explicable on the basis of differences in I.Q. accuracy for I.Q.'s of differing magnitudes; i.e. it is no longer necessary to postulate such explanations as a possibly greater irregu-

[1] N.D.M. Hirsch, "An Experimental Study of Three Hundred School Children over a Six-Year Period," *Genet. Psychol. Monogr.*, 1930, 7, 487-547.

larity of growth for the superior or a supposed differential operation of motivational or emotional factors.

So far our discussion has pertained primarily to the errors of measurement for I.Q.'s even though the real source of error is, of course, in the M.A. scores, but knowing either error we can secure an approximate value for the other by use of the general relationship $\sigma_{e(MA)} = (CA)\sigma_{e(IQ)}$. This would need to be modified for individuals of 14 years and over because of the substitute C.A. divisor. The standard errors of measurement for M.A. scores, as given in Table 22, have been ascertained, however, by two schemes which will be described briefly. By the first method, we obtained from the separate age scatter plots, for I.Q. differences versus magnitude of I.Q., a regression estimate of the difference in I.Q. for those who score at age, i.e. have I.Q.'s equal to 100. This difference in I.Q.'s times the C.A. gives the corresponding difference in M.A. score, and from this last difference we get an estimate of the error for an M.A. score by multiplying by $\frac{1}{2}\sqrt{\pi}$. The second method involved plotting the actual differences between the mental-age scores derived from the two forms against mental age, then computing an average difference for the several M. A. levels, 30-39, 40-49, etc., and again multiplying by the appropriate constant to obtain $\sigma_{e(MA)}$. A freehand or graphic method of smoothing the results from these two methods yielded the values in Table 22. We believe the values so obtained to be satisfactory approximations for the errors for M.A.'s below 180 months. It would seem safe to accept 7.5 as a fair estimate of the standard error of measurement for M.A.'s higher than 190 months, but we are not sure some other value between 7.0 and 8.0 is not just as correct.

It should be noted that the values in Table 22 are for the particular M.A.'s given and not for a group of values; e.g. 2.9 is the error for an M.A. score of 60 months. Linear interpolation can be used for in-between values. Those who attempt to check back and forth from

TABLE 22

STANDARD ERROR OF MEASUREMENT IN MONTHS FOR MENTAL AGE SCORES (APPROXIMATE)

M.A.	$\sigma_{e(MA)}$	M.A.	$\sigma_{e(MA)}$
180	7.0	100	4.6
170	6.5	90	4.2
160	6.2	80	3.9
150	6.0	70	3.5
140	5.8	60	2.9
130	5.5	50	2.5
120	5.2	40	2.2
110	4.9	30	1.9

$\sigma_{e(IQ)}$ to $\sigma_{e(MA)}$ will find apparent discrepancies which are to be explained on the basis of the values for I.Q.'s having been obtained from age combinations with no allowance for changes in reliability of the scales within an age combination; whereas the method of estimating the errors for M.A.'s will actually reflect such changes in reliability with age.

In concluding this chapter, it should be noted that the measurement errors reported herein have not been adjusted for duplicate items in the two forms. Mention should also be made of the fact that the standardization testing involved administering more items than in the present forms; hence, more than ordinary time was required, with the possible result that the reliabilities have been somewhat lowered by fatigue or motivational factors. A compensating factor exists in that the testing was done under optimum conditions by well-trained examiners. Ordinary routine testing may not lead to as high reliability as our coefficients would indicate.

Chapter VII

SPREAD OF INDIVIDUAL PERFORMANCES[1]

During the standardization-testing period, it was noticed that the spread of performance, i.e. the failing of tests below and the passing of tests above an individual's general mental level, seemed to show greater variation than that generally observed for the 1916 Revision. It was thought that this greater spread might, at least in part, be due to the fact that certain test items were poorly located in the preliminary forms, but subsequent use of the final forms, both at Stanford and elsewhere, has revealed what seems to be a rather wide scatter of passes and failures. It is the purpose of this chapter to discuss, and to set forth some possible explanations for, this variability, and to present the results of an attempt to analyze the standardization data in such a manner as to check on certain hypotheses regarding its cause.

That the performance of nearly all individuals must show some scattering of passes and failures is a foregone conclusion. This is evident when we recall that the scales are made up of items which yield per cent passing with age curves which are none too steep and items which are not perfectly intercorrelated. Insofar as items may measure a group function in addition to the general and specific abilities, and insofar as these abilities, group and specific, may develop at different rates within the individual, it follows that an individual is apt to fail items below or pass items above his general level of performance. Another obvious reason for variability is the relative unreliability of single items. It should also be noted that the spread of performance or

[1]The investigation reported herewith was financed in part by a grant from the Committee on Psychology and Anthropology of the National Research Council.

variability will increase from the lower to the higher age levels. Whether this increase in absolute spread is more apparent than real can only be answered when a satisfactory answer is given to the larger problem to which this is related, i.e. the problem of increasing individual differences with age. The fact that the curves of per cent passing with age, though steep at the earlier ages, tend gradually to flatten as one approaches maturity ages, is sufficient cause for an increase in apparent individual variability with age.

But these factors which we have just enumerated are not necessarily valid reasons for the observed greater scatter on the new forms as compared with the 1916 scale. One possible explanation for the greater variation is an artifact of certain differences between the old and new scales. The presence in the new scales of test items at age levels 11 and 13 and the additional items at the upper age levels is the one definite reason for an apparently greater spread for individuals of ages 9 to 15. Likewise the presence of tests at half-age levels at the lower end of the scale will result in an apparently greater spread. It cannot be argued that the larger scatter, if it really is larger, on Forms L and M is due to faulty placement of items; nor can it be said that it is due to the inclusion of items which show a wider age range for passes and failures than the items in the 1916 Revision. Aside from the aforementioned location of items at additional age levels, we see no real reason for a greater spread of performance on the New Revision. The presence of spread, however, does call for study, so we are here reporting an attempt, none too fruitful, to investigate two aspects of the problem.

First we have sought an answer in terms of the nature of the items to the question: Is the spread of passing and failing a function of the scale itself in that certain items, or sequences of items, tend to measure abilities other than the general ability demanded by all items? This cannot be answered by recourse to the re-

sults of the factor analyses, since those analyses were necessarily confined to contiguous items. In a second line of inquiry we have set out to learn whether or not the spread or variability of performance is a function of the individual. Here we are granting that the instrument is such as to permit variability of performance, and we are raising the question as to whether individuals are consistently variable or nonvariable. Later it will be seen that the answer to this last question will need to be modified in light of the answer given for the first question.

Spread as a Function of the Scale

The question regarding the spread of passing and failing as a function of items in the scale can be subdivided into more specific questions. To what extent is the spread due to items which recur with different passing standards at different ages, to items which are highly similar, such as repeating digits and other recurring test situations, and to items which seem to have something in common over and above the general factor? The most satisfactory answer to the last two parts of this question could be determined by factor analyses far more extensive than is feasible with so few as 100 to 200 cases at an age level. In fact it would require several thousand cases at each level to compensate for the error involved in extreme dichotomies (for passing and failing) for those items four or five age levels removed from the given age.

Our approach has been by way of an examination of the patterns of failing (and passing) for those individuals whose spread seems abnormal. In order to make such an analysis, the spread for all cases on each form was tabulated, separately for each age level, in such a manner as to yield distributions which showed the extent of passing above, and failing below, each individual's mental-age level. This was not done for ages below 4

and above 15, since the spread downward and upward respectively for levels beyond these is automatically curtailed. On the basis of these distributions it was possible to pick out those cases with spread of passes extremely above their mental age level and those cases which showed extreme downward variability for failures below their M.A. level. An individual might be included in both extremes - the extent to which spread upward is correlated with spread downward will be reported later in this chapter. The definition of 'extreme' spread is arbitrary - in order to have sufficient numbers, about 28 per cent of the cases have been designated as extreme. It is of course very doubtful whether so large a percentage can be considered as really representing unusual variability.

A record was made of the particular items passed at all levels above an individual's M.A. level for each of the 709 subjects chosen as showing extreme passing spread, and a similar record was made of the failures below M.A. level for each of 757 extreme cases of downward spread. Then a tabulation was made of the patterns of items failed by those with downward spread of failures, and separately for the patterns of successes for those with upward variation of passes. In determining the frequency of occurrence for the patterns, the data were grouped according to M.A. levels, since individuals of the same M.A. level have more trials or attempts in common than those of the same C.A. In so arriving at patterns and their frequencies, the results for Forms L and M were kept distinct. As one might expect, the total number of patterns which occurred was very great since the possible number of combinations is large. Some patterns occurred more frequently than others, but there seems to be no satisfactory statistical method for setting up criteria as to how large frequencies should be in order to be considered of statistical significance. Such criteria would involve knowing the probability for the occurrence of patterns which are spread over two to six

levels, but the solution of this problem would involve empirical probabilities for the passing (or failing) of single items by an individual of a specified M.A. Even if this could be resolved, the chi square comparison of observed frequencies with expected frequencies, so deduced, would be highly questionable because of the very small expected frequencies.

It should be noted that the tabulations yielded the frequencies for the patterns of failures (or passes) as they occurred two, three, four, five, or as many as six levels below an individual's M.A. level (above for passes). Thus the record of the items failed (or passed) adjacent to the given M.A. level could be examined in order better to interpret the patterns of items failed (or passed) at levels remote from an individual's mental age. For example, 28 of the 110 individuals having mental ages from 11 to 12 on Form L passed test S.A.I, 3. An examination of the other items passed at this extreme level and the two adjacent levels, A.A. and 14, by these individuals does not lead to a plausible explanation for this spread upward of passing behavior, but when the items which were passed nearer the 11-year level were scrutinized, a definite reason was found for this extreme passing performance: item S.A.I, 3, 'Minkus completion,' recurs from age level 12. Any individual who happens to possess the specific ability called for by this test and who gets three sentences completed, automatically receives credit for this test at the higher level. This example, which is unusual in terms of frequency, is in line with one of our original hypotheses; namely, that a part of the spread is due to the recurring tests.

The presentation here of the tabular and numerical data which accumulated during this quest is certainly not feasible, hardly necessary, and perhaps unwarranted when one considers the complexity of the data and their inappropriateness to statistical treatment. We shall, therefore, present such results and conclusions as seem to emerge rather clearly when the patterns are carefully

examined. It is admitted that this may not be very satisfactory to the reader, but we feel certain that the data do not justify more elaboration.

Let us first consider the passes for those showing an extreme upward spread from their M.A. levels. On Form L, in addition to the 'Minkus completion' test cited above, we find that a few patterns involve 'reconciliation of opposites,' a test located at level A.A. and recurring at S.A. II. Aside from these two recurring tests plus 'abstract words,' which has a negligible frequency, one must search elsewhere for an explanation for most of the spread of passing on Form L for mental levels four to average adult. The recurring 'vocabulary' test and 'repeating digits' (recurring test situation) are both conspicuous by absence from the patterns. But since the 'vocabulary test' is highly saturated with the general function measured by the entire scale, one would not expect, on the basis of our hypothesis, to find that it contributed to variability of performance. Its specific part (factor) is indeed small. On Form M, four recurring tests ('picture vocabulary,' 'orientation: direction,' 'reconciliation of opposites,' and 'ingenuity') would seem to have produced a part of the extreme passing performance. The 'Minkus completion' test of Form M is not involved (a marked inconsistency with the finding for Form L), while 'repeating digits' and 'picture absurdities' as recurring test situations occur in the patterns with negligible frequencies.

When we turn to failures below mental age, we first note that the spread downward is greater than the upward spread of passing, and consequently the number of patterns and possible patterns is much greater for failures below than for passes above a given mental level. For Form L, the following recurring tests occurred, rather frequently in the patterns: 'picture vocabulary,' 'three-hole form board: rotated,' 'pictorial identification,' 'picture completion: man,' 'vocabulary,' 'identifying objects by use,' 'memory for designs,' and 'paper cutting.' On the other form four recurring tests were found to be involved:

'identifying objects by use,' 'patience: pictures,' 'verbal absurdities,' and 'abstract words.' These tests appeared in only a fraction of the total number of patterns, but those patterns containing them (except 'vocabulary') occurred much more frequently than other patterns; i.e. a rather large number of individuals was affected thereby.

Five items which recur not as the same test but as a similar test situation ('picture absurdities,' 'comprehension,' 'verbal absurdities,' 'digits,' and 'memory for sentences') were found to be possible contributors to failing spread on Form L; and on Form M the following test situations were involved: 'verbal absurdities,' 'memory for sentences,' 'abstract words,' 'memory for designs,' and 'picture absurdities.' These recurring test situations do not automatically produce a failure, as does a poor performance on an actual recurring test, but in case an individual is somewhat lacking in the specific ability needed and in case he meets the situation somewhat below his general level of performance, he will be apt to fail and thereby add to his variability score.

It is thus seen that a part of the spread of performance is definitely linked to recurring tests and recurring test situations, but it cannot be claimed that this is an explanation for anything like all of the extreme variation. There may, however, be an indirect connection between failing to meet the criterion for passing a recurring test and the failing of other items at the lower level to which an individual is thereupon taken back. Some of these easier items may not be of sufficient challenge to him. It is interesting to note that only about one-half the recurring tests and recurring test situations which are included in Forms L and M were involved in the patterns which we have examined. The reason for this seems obscure.

SPREAD OF INDIVIDUAL PERFORMANCES

Spread as a Function of the Individual

That the variability of performance on examinations of the Binet type is a characteristic of the individual has been frequently postulated, and Kuhlmann in his 1939 *Tests of Mental Development* presents norms for variability scores. He does not, however, claim any significance for his measure of variability, but rather gives norms so that if, and when, psychological or clinical meaning can be attached to a variability score, they will be available. One of the first questions to be raised about a variability score is its stability from test to test; i.e. does an individual show consistently high or low variability about his general level of performance? In answering this question, we shall accept the usual variability score as the distance, in terms of age levels, from an individual's basal mental age to the highest level at which tests are passed.

Aside from the consistency of variability scores, there is another aspect of variation which may be of some interest, namely the relationship of upward spread of passes to downward spread of failures. Let us first present some data on this point. The correlation (tetrachoric because of fewness of categories) of upward versus downward variation was determined for three mental-age groupings, the groupings being made separately for, and on the basis of mental ages on, Forms L and M. Although for a particular mental-age group, the cases for the Form L correlation will not be exclusive of those involved in the correlation for Form M, the cases will not be exactly the same and the N's need not agree. These correlational results are given in Table 23, from which it can be seen that there is only a slight relationship between upward and downward (from an individual's M.A.) spread of performance. Insofar as passing above reflects high motivation and failing below represents poor effort, the lack of correlation is not surprising. Variability of

TABLE 23

CORRELATION BETWEEN UPWARD AND DOWNWARD SPREAD OF PERFORMANCE FOR CERTAIN LEVELS

Mental age	54-59		120-131		144-155	
Form	L	M	L	M	L	M
N	113	122	204	201	201	198
r_t	.20	.18	.17	.06	.27	-.03
σ_r (r=0)	.18	.15	.11	.11	.11	.11

motivational factors for an individual during the test administration would, however, tend to produce correlation. The above correlations include a small spurious or artifactual element in that both the upward and downward spreads are measured from the individual's M.A. level, which in a sense is an average of the two variations. For instance, it is unlikely that a high upward spread could be accompanied by no failures below one's mental-age level.

In order to obtain some information on the consistency of variability scores, we have correlated the variation score as determined from Form L with that secured on Form M for six groups at different levels of maturity. As in the case of the correlations just reported, mental-age groups were used instead of life-age groups because variability of performance as measured is not correlated with C.A. for constant mental age whereas it is related to M.A. when C.A. is held constant. Mental age is in this case based upon the composite of Forms L and M.

In reporting the resulting product-moment correlations and means and standard deviations (see Table 24), we have not made any correction for two factors which disturb these values for the three lower groups. These are (1) the fewness or coarseness of the categories and (2) the counting, in determining the variability score, of spread over half-age and full-age levels as though they

were of the same value. Since the tetrachoric r's, which will not be much affected by either of these factors, are in close agreement with the reported product-moment coefficients, we assume that the error introduced is not such as to disturb seriously our general conclusion to the effect that these data lead one to question the consistency or reliability of individual variability scores. At

TABLE 24

CORRELATION FOR VARIABILITY SCORE ON FORM L WITH THAT ON FORM M

M.A. group	42-47	60-65	84-95	108-119	120-131	144-155
N	95	125	187	174	204	209
M_L	4.15	3.36	4.32	6.07	6.28	6.66
M_M	4.48	3.96	4.26	6.06	6.31	6.78
σ_L	1.02	1.35	1.41	1.57	1.46	1.52
σ_M	1.04	1.11	1.60	1.35	1.44	1.63
r	-.077	.304	.367	.360	.190	.059

this time, it should be pointed out that insofar as spread of performance is a function of the nature of the scales, particularly the presence of highly similar tests (such as recurring test situations) which introduce or permit narrow group factors, and in case the same narrow factors are involved in both forms, we have a condition which would tend to produce some correlation between variability on the two forms. Hence, we may conclude that these correlations represent a sort of upper limit for the 'reliability' of individual variability scores. If this be true, and unless these correlations are affected by unknowns, we are forced to conclude that little significance can be attached to measures of this type of individual variation.

The statement made above to the effect that variability is not correlated with C.A. for constant mental

age implies, of course, that variability is not related to brightness when mental maturity is kept constant. In order to check on this point, four different M.A. groups were chosen and the correlation between variability and I.Q. was determined. This was done for Form L only. The results are given in Table 25, from which we infer that the relationship is possibly negative, but the r's are so near zero that very little of the variability variance can be attributed to, or associated with, brightness.

TABLE 25

CORRELATION BETWEEN VARIABILITY AND BRIGHTNESS OR I.Q. FOR CONSTANT M.A. GROUPS. FORM L

M.A. group	72-83	96-107	120-131	144-155
N	263	196	204	201
r	-.097	-.016	-.097	-.098
σ_r (r=0)	.062	.071	.070	.071

In closing this chapter, a few limitations of the data analyzed herein should be mentioned. The standardization-testing involved preliminary forms which contained more items than the final forms, a different order for some of the items, and in some cases rather poorly placed items as regards difficulty. It is not possible to say how much these factors would invalidate such conclusions as have been drawn; or, to put it differently, we cannot be positively sure that the results are the same as would be found if similar analyses were made on new data based upon examinations with the final forms. Our conjecture is that substantially the same findings would emerge.

Chapter VIII

PER CENTS PASSING ITEMS BY AGE

In constructing an age scale, the per cent passing an item at successive age levels takes on considerable importance. In the first place, such percentages supply some necessary, though not sufficient, information for establishing the validity of an item. In the second place, they provide the necessary data for arranging items according to difficulty, i.e. locating the items in the scale. These two main points are so well known that one should not need to discuss them, but since some people persist in misunderstanding the role played by curves of per cent passing by age, it may not be amiss to recapitulate briefly the rationale underlying their use.

To determine the validity of an intelligence-test item is a far greater task than to ascertain its difficulty. One begins either with a common-sense notion of intelligence, usually delimited as to kind, or with a high-powered definition couched in the jargon of contemporary psychology. Then one searches for items which will, according to the best judgment, provide behavioral situations calling for the kind of intelligence implied in the definition. That a particular psychologist's definition of intelligence is not the main determiner of the outcome is attested to by the fact that the tests constructed by individuals having different conceptions of intellect tend to be highly intercorrelated. This, of course, does not prove that the items selected a priori by the several test-makers are valid; it merely demonstrates that there is some unanimity as to what is to be measured and as to the general type of situations which will reflect individual differences in intellect. The test-constructor must use all possible adjuncts for selecting items so as to produce a test which satisfies the ultimate criterion of social utility.

Now, there would seem to be agreement that general intelligence is a characteristic which develops with age, so it is entirely logical to lay down the requirement that an item cannot be regarded as valid unless it yields a larger per cent passing for successive age levels through childhood. It should go without saying, however, that this requirement in and of itself does not guarantee validity. For instance, the ability to throw a baseball a distance of twenty-five feet will show an increase in success with age, but such an item would, for obvious reasons, never reach the tryout stage. All test-constructors have recognized the necessity of other criteria of validity in addition to increase in per cents passing at successive age levels. Among the additional criteria most frequently used are correlations with subjective ratings of intelligence, correlations with scores earned on other intelligence scales, and correlations with total score on the battery of which the item is a part. The merits and limitations of these and other criteria have been discussed in Chapter I. As there stated, the elimination of test items from the trial series for the new Stanford-Binet revision was based in part on correlation with the composite of L and M scores, a procedure which insures that retained items will be saturated with a common factor.

With regard to the arrangement of items according to difficulty, and their allocation to a given age level, it should first be remarked that no attempt was made to arrange the items within an age level in order of difficulty. This, despite the carping of a few critics, cannot be regarded as a serious drawback, since the differences in difficulty within a level are usually small. Besides, there are times when psychological reasons should take precedence over purely statistical dictates, particularly when the statistical differences are of negligible practical significance.

Many have pointed out the difficulties in constructing an age scale, but only those who have been through

the mill are in a position to appreciate fully the obstacles. At times, compromises must be made - all contingencies cannot be foreseen. It happens that a few items were retained which were not entirely satisfactory as regards their final allocation, or passing curves. For instance, it will be seen from the table of per cents passing that item M, III, 4 is easier than M, II-6, 3, but the two could not be switched because the latter item was needed at level III, with a higher passing standard. Item L, V, 3 would appear to be much easier for ages 3-1/2 and 4 than the other items at level V; and M, VII, 2 is somewhat more difficult for age 8 and up than the other items located at level VII. Items L, XIV, 2 and M, XIII, 2 yield curves which tend to flatten too much (i.e. more than desirable) for the four upper age groups. Test L, XI, 1 would seem to be the worst misplaced of all, as it corresponds in difficulty to items located at level XIII.

A number of clinicians have expressed the belief that many items are not in their proper order as regards difficulty. This may be true for selected or special groups; but before one can generalize one must have actual data rather than impressions, and these data must be based on representative samplings. Furthermore, the size of the age samplings must be such as to make for some stability in the percentages for passing. As an example of a decidedly inadequate study with unwarranted conclusions regarding test placement, we cite the paper of Growdon.[1] The order of difficulty of test items is not something that can be established once and for all. It can be expected to vary somewhat with samplings of different populations, especially samplings from different

[1] C. H. Growdon, "Is the Revised Stanford-Binet Scale Really an Age Scale?" *Psychol. Bull.*, 1940, 37, 512. (Abstract: the writer heard this paper at the Pennsylvania State College A.P.A. meeting.)

countries. It is not surprising, for example, that Burt[1] finds the order of difficulty of the New Revision items for London children somewhat different from that for the American standardization group. One is a little puzzled, however, as to the meaning of Burt's assertion that 'there seems to be no fixed order at all. what is easier for one child may be harder for another.'

We would also call attention to the fact that the curves of per cents passing by age can affect the variability of I.Q.'s. The reader may recall the discussion in *Measuring Intelligence* (page 40) concerning the rather marked fluctuation in standard deviations of I.Q.'s for various age groups. In particular it was noted that the standard deviations were decidedly too low for age 6, and somewhat high for ages 2-1/2 and 12. At the 1937 writing no explanation had been found for these apparent facts, but it was pointed out that there was nothing in the samplings to suggest that the atypical standard deviations might be due to selective factors. Since then, a closer scrutiny of the curves for per cents passing has convinced us that the difference in variability is an artifact of the scale. It results from strange, and undetected, accidents. It is well known that the extent of variability is partly a function of item difficulty. It so happens that no items yield curves for per cents passing which cross the ordinate for age 6 between the 35 and 65 per cent levels of difficulty. This fact will definitely result in a narrower spread of M.A.'s and I.Q.'s for 6-year-olds than would have resulted had this imperfection been absent. The same situation as regards lack of items of medium difficulty exists for ages 5 and 5-1/2, though not so markedly as for age 6. The greater variability at ages 11 and 12 may be due to a concentration of items of medium difficulty for these ages, a concentration which is greater than that at other ages, except at

[1] Cyril Burt, "The Latest Revision of the Binet Intelligence Tests," *Eugen. Rev.*, 1939, 30, 255-260.

2-1/2, 3, and 3-1/2. There is, however, nothing about the item difficulties for age 15 which would enable one to predict the rather large S.D.'s for I.Q.'s at that age. Although it is unfortunate that these differences in variability should exist as an artifact of scale construction, it is somewhat satisfying to know that one need not entertain the psychoanalytic explanation offered by Bellak.[1]

Anyone who examines the data of Table 26 for per cents passing will note that the plan of the 1916 revision has again been used in allocating items to the several age levels. This procedure involves a shifting standard as to difficulty as one proceeds from the lower to higher age levels. For instance, the items located at level II are of such difficulty that they are passed by about 77 per cent of 2-year-olds; the level V items are passed by about 70 per cent of 5-year-olds; those at VIII by about 63 per cent of 8-year-olds, etc. The large jump in difficulties at the adult levels is, of course, necessary in order to provide 'top,' but our chief interest just now concerns the relationship between difficulty and item placement.

It is quite evident that there is considerable misunderstanding of the issues involved here. A very unusual misconception comes from the pen of M.W. Richardson.[2] It might have been expected that a critic of Binet method would have better informed himself about age scales before making the statement that 'the age at which just half the children pass the test is taken as the scale-position of the item.' Some five pages later Richardson goes on record as believing that subtests are properly scaled only when assigned to the year level yielding 50 per cent passing. Guilford[3] evidently holds

[1]See L.Bellak, "A Possible Dynamic Explanation of Variability in the I.Q.," *J. Abnorm. (Soc.) Psychol.*, 1941, 36, 106-109.

[2]M. W. Richardson, "The Logic of Age Scales," *Educ. Psychol. Measmt.*, 1941, 1, 25-34, especially p. 26.

[3]J.P. Guilford, *Psychometric Methods*. New York: McGraw-Hill Book Company, 1936; see p. 409.

to the same view. He states that 'a proportion of 50 per cent would have been a better criterion of age level. A knowledge of psychophysical procedure, as in the constant methods, would have suggested this criterion.'

It seems to us that the misunderstanding about the Binet method of a sliding scale of difficulty for allocating items to age levels may be due to failure to appreciate any one or more of the following considerations. (1) The fact that it is simply impossible otherwise to construct an age scale of the Binet type that will yield mean mental ages equal to mean chronological ages. (2) The fact that the location and grouping of items at a given level is mainly one of convenience which facilitates testing and scoring. This convenient way of arranging tests is of course closely related to the first consideration. (3) The fact that the individuals of any age group encounter items which are actually of 50 per cent difficulty for their age group even though the items placed at their own age level may be less difficult. The criticism for not 'properly' locating items is actually invalidated by this third point.

The presentation of Table 26 for per cents passing each item by age will no doubt become an invitation to some to use these data for establishing a growth curve. Perhaps a new 'absolute' zero point for intelligence may be found, and no doubt such a curve (or curves if the items for Forms L and M are treated separately) will shed apparent light on the question as to when mental maturity is reached. These data, however, may not be entirely satisfactory for studying mental growth. As in all cross-sectional studies, one must be reasonably sure that the several age samples are strictly comparable as regards their representativeness. Any selective factors present will tend to distort the derived growth curve, and such distortion can become a serious restriction to extrapolation. It happens that our preschool samples and those for the top two or three ages are not representative. This fact led to an intentional

adjustment in the direction of permitting mean I.Q.'s above 100 for children of our sample at these levels. Whether or not the allowance has been adequate will become more or less evident with use. However, the pertinent point to note here is that the per cents for passing have not been adjusted for known inadequacies in the sampling procedure. This fact is, we believe, sufficient to render highly questionable the meaning of growth curves which might be derived from these data by any of the so-called scaling methods.

TABLE 26

PER CENTS PASSING ITEMS BY AGE

(Names of tests may be found in Appendix C)

Item	Age 1½	2	2½	3	3½	4	4½	5
L,II,1*	18	79	97	99	98	98	100	
2	37	86	95	97	99	95	99	
3	11	74	92	92	99	94	99	
4	23	67	85	92	97	99	100	
5	13	79	92	94	99	98	100	
6*	28	77	98	96	100	98	99	
a	28	64	84	93	94	97	98	
M,II,1	50	80	83	90	96	99	100	
2	39	81	98	98	99	98	100	
3	10	68	97	93	99	98	99	
4*	18	79	97	99	98	98	100	
5	14	72	92	93	99	98	100	
6*	28	77	98	96	100	98	99	
a	0	61	88	90	99	99	99	
L,II-6,1	10	52	77	89	95	97	98	100
2	5	58	82	88	97	97	99	100
3	0	24	76	84	99	98	99	99
4	0	23	77	88	97	98	98	100
5	0	38	74	86	92	98	97	99
6	5	45	72	79	92	95	97	99
a	18	65	87	91	98	98	99	100
M,II-6,1	2	26	76	86	96	98	98	100
2	10	45	80	92	97	99	100	100
3	0	26	64	81	96	97	98	99
4	0	32	76	89	99	98	99	100
5	2	38	72	83	97	98	96	100
6	7	45	68	84	94	97	98	100
a	2	17	77	92	98	99	100	100

*Item duplicated on other form.

TABLE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age 1½	2	2½	3	3½	4	4½	5	5½	6
L,III,1	0	9	50	73	95	96	100	100	100	
2	0	11	51	73	94	97	97	96	100	
3*	0	5	36	73	89	95	99	100	100	
4	6	27	52	65	82	86	97	98	100	
5	0	12	30	62	84	94	98	98	99	
6	0	15	55	76	89	91	95	98	99	
a*	2	23	49	65	80	89	90	98	99	
M,III,1*	0	5	36	73	89	95	99	100	100	
2	0	7	44	69	92	97	97	100	96	
3	0	13	51	67	89	94	96	98	99	
4	6	27	63	87	95	96	98	100	100	
5	0	6	42	69	86	95	97	99	99	
6	0	8	38	64	83	91	95	97	99	
a*	2	23	49	65	80	89	90	98	99	
L,III-6,1		10	48	62	74	82	91	96	98	96
2		1	13	32	64	78	84	92	96	99
3		1	14	41	67	82	91	92	94	98
4		3	27	52	78	90	93	99	99	99
5		11	42	62	74	87	95	96	97	100
6		6	22	38	72	79	94	98	98	99
a		0	14	43	70	78	93	96	98	100
M,III-6,1		1	19	58	79	85	86	91	93	98
2		2	9	38	66	71	87	96	97	99
3		0	8	29	69	86	91	98	100	99
4		2	17	39	75	82	78	89	88	94
5		0	22	54	81	90	94	97	100	99
6		5	34	49	79	92	95	99	100	100
a		1	9	29	65	75	90	96	99	100

*Item duplicated on other form.

TABLE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age 2	2½	3	3½	4	4½	5	5½	6	7	8
L,IV,1	0	7	17	41	61	72	86	84	95	97	
2	1	15	25	63	79	76	89	94	98	100	
3	0	12	41	57	74	84	88	94	95	98	
4	0	11	29	62	77	83	93	97	97	98	
5	4	5	17	40	64	80	88	94	96	99	
6	0	9	14	57	62	75	88	93	97	100	
a	0	12	25	70	73	84	90	92	96	100	
M,IV,1	2	20	40	69	82	91	94	100	99	100	
2	0	7	28	65	77	95	100	99	99	100	
3	0	4	8	38	62	79	88	92	97	98	
4	1	17	28	66	72	78	91	94	98	100	
5	3	15	32	58	72	89	97	100	99	99	
6	2	22	37	70	81	91	97	95	100	99	
a	0	6	13	49	70	80	88	96	98	100	
L,IV-6,1	0	6	21	40	55	76	79	90	96	98	100
2	1	12	18	49	47	74	77	84	85	97	100
3	1	13	27	35	53	71	82	84	94	100	99
4	0	1	8	30	38	66	77	82	90	97	99
5	0	2	14	38	53	71	75	78	90	97	100
6	0	1	3	33	48	67	78	78	93	99	100
a	0	5	11	33	43	66	77	89	97	98	100
M,IV-6,1	0	5	10	38	56	73	83	94	96	99	100
2	1	10	24	56	67	73	87	96	94	99	100
3	0	9	13	39	42	64	78	79	89	98	99
4	0	3	16	37	58	72	81	91	94	98	100
5	0	1	6	40	59	79	85	94	96	100	100
6	0	12	13	50	59	77	86	88	96	98	100
a	0	3	20	52	60	78	94	96	97	100	99

TABLE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age										
	2½	3	3½	4	4½	5	5½	6	7	8	9
L,V,1	3	6	19	27	50	66	75	86	97	100	
2	1	2	19	38	65	82	85	92	97	100	
3	11	18	49	65	66	86	94	94	96	98	
4	0	0	4	16	50	67	82	95	99	100	
5	2	2	24	27	52	64	76	83	90	94	
6	2	3	22	30	50	68	81	91	99	99	
a*	0	3	6	20	44	69	74	92	98	100	
M,V,1	6	15	33	46	66	78	92	93	98	100	
2	3	10	24	39	50	79	87	96	99	100	
3	0	0	13	25	51	70	72	84	98	100	
4	1	4	16	33	51	63	63	77	92	98	
5	0	2	18	22	43	65	76	90	96	99	
6	0	0	7	18	41	69	76	85	89	97	
a*	0	3	6	20	44	69	74	92	98	100	
L,VI,1			0	3	15	36	50	67	89	97	99
2			0	11	29	44	55	70	86	95	99
3			5	11	26	46	53	69	86	96	98
4			1	3	11	43	48	71	94	96	99
5			7	16	29	47	51	73	94	95	100
6			20	26	44	52	61	81	91	93	99
M,VI,1			0	3	16	53	60	75	96	99	100
2			0	5	25	42	64	81	91	97	99
3			0	8	19	40	46	68	90	93	98
4			0	18	40	62	74	83	96	98	100
5			2	5	20	39	54	73	95	100	100
6			10	18	29	50	50	75	85	96	98

*Item duplicated on other form.

TABLE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age 4	4½	5	5½	6	7	8	9	10	11	12	13	14	15
L,VII,1	0	5	5	28	39	59	78	85	96	98	99	100		
2	4	9	12	17	28	51	74	84	93	97	95	100		
3	0	1	3	10	23	69	91	95	96	96	99	100		
4	0	2	12	17	31	59	78	82	92	93	94	95		
5	1	3	10	16	20	55	74	84	93	95	96	98		
6	8	22	25	35	41	70	81	87	96	96	95	96		
M,VII,1	0	1	12	14	36	81	96	100	99	100	100	100		
2	0	3	2	11	24	55	63	74	86	87	92	94		
3	1	5	10	8	28	55	80	88	94	97	96	100		
4	0	2	13	20	34	72	86	93	96	98	100	100		
5	0	5	5	12	18	50	74	84	94	97	98	98		
6	0	5	7	15	36	67	82	95	93	96	99	100		
L,VIII,1			4	10	17	44	66	83	93	98	97	99	100	100
2			0	5	14	32	67	82	90	87	90	95	94	93
3			1	4	8	30	58	81	82	91	90	96	97	98
4			3	1	8	31	57	74	85	90	92	96	97	97
5			1	6	10	23	58	75	85	91	93	95	97	98
6			7	14	23	44	59	74	75	84	87	95	93	95
M,VIII,1			2	10	21	45	64	76	83	94	94	98	98	100
2			8	9	19	40	69	85	92	95	95	98	98	99
3			1	6	7	27	58	76	85	91	92	98	98	96
4			4	1	12	43	70	84	94	97	98	100	100	100
5			2	12	17	42	62	74	86	94	94	98	99	96
6			7	9	17	39	67	76	91	93	95	97	98	96

TABE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age													
	5½	6	7	8	9	10	11	12	13	14	15	16	17	18
L,IX,1	0	4	21	42	64	68	77	82	90	92	90	94	95	
2	1	1	7	27	48	67	79	77	89	90	92	95	99	
3	1	2	22	31	50	60	72	76	91	92	88	91	94	
4	1	4	14	32	56	70	76	83	92	92	91	92	99	
5	0	0	11	39	70	84	94	92	95	100	97	98	99	
6	3	6	26	38	61	79	79	86	94	94	95	95	99	
M,IX,1	3	4	17	31	56	64	76	84	92	93	90	94	99	
2	1	0	8	36	67	82	87	91	96	97	98	99	100	
3	1	2	9	27	52	69	83	75	88	90	95	96	98	
4	1	4	14	36	56	61	70	77	89	92	91	94	94	
5	2	5	21	40	59	71	69	80	82	80	84	91	93	
6	4	9	28	49	69	84	80	90	96	95	96	96	98	
L,X,1		0	5	15	39	59	76	86	92	96	99	97	100	98
2		2	10	19	40	58	61	70	74	72	77	77	88	84
3		4	2	10	29	56	67	61	75	71	74	79	80	76
4		0	10	25	46	63	60	74	83	89	90	92	99	94
5		0	9	28	40	55	66	70	83	86	88	94	98	89
6		9	19	35	48	67	63	76	80	87	88	92	91	93
M,X,1		5	16	28	40	54	62	71	76	84	84	89	91	96
2		0	5	26	46	62	80	89	91	92	93	97	98	94
3		1	5	20	37	54	66	66	74	69	77	81	94	92
4		0	5	18	39	60	77	84	93	90	97	97	99	97
5		10	24	38	52	67	77	79	93	90	93	96	90	98
6		10	27	42	52	65	65	73	81	82	89	94	96	92

TABLE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age 7	8	9	10	11	12	13	14	15	16	17	18
L,XI,1	6	9	21	30	43	57	67	71	76	86	85	90
2	7	16	34	51	65	75	80	86	91	86	96	91
3*	0	2	14	28	58	71	84	84	89	91	97	96
4*	4	14	30	40	50	61	74	72	80	85	87	88
5	10	22	40	52	65	74	83	84	86	92	95	94
6*	4	17	34	48	59	72	73	81	79	85	92	89
M,XI,1	5	18	29	44	54	59	74	81	87	87	94	96
2	7	24	36	51	64	69	72	80	81	81	86	89
3	2	11	33	48	68	60	76	77	87	94	97	93
4*	0	2	14	28	58	71	84	84	89	91	97	96
5*	4	17	34	48	59	72	73	81	79	85	92	89
6*	4	14	30	40	50	61	74	72	80	85	87	88
L,XII,1	1	3	10	21	46	64	76	81	89	95	98	96
2	4	11	24	47	56	62	75	85	85	89	97	93
3*	2	10	26	43	56	69	72	70	74	80	82	79
4	4	9	20	39	48	63	77	71	75	80	82	83
5	2	2	10	27	52	68	75	80	80	79	89	78
6	3	11	26	40	52	61	77	72	73	80	81	87
M,XII,1	1	7	28	37	60	65	72	81	83	84	90	91
2*	2	10	26	43	56	69	72	70	74	80	82	79
3	4	6	21	35	46	61	69	71	77	78	89	89
4	1	4	15	33	54	68	78	80	90	91	95	94
5	6	11	33	37	53	62	72	71	77	83	88	85
6	3	13	33	47	51	59	72	77	72	75	85	88

*Item duplicated on other form.

TABLE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age 7	8	9	10	11	12	13	14	15	16	17	18
L,XIII,1	13	21	29	39	41	53	60	65	63	76	74	75
2	4	8	22	38	51	58	69	69	72	80	82	89
3	0	5	20	28	41	47	60	62	70	77	81	79
4*	1	6	25	40	52	63	69	74	74	81	82	74
5*	0	2	10	20	44	49	66	69	71	79	84	90
6	5	10	24	35	52	59	64	70	76	72	91	83
M,XIII,1	13	22	27	38	44	47	54	66	64	80	74	73
2	1	5	13	29	45	49	59	53	59	57	54	56
3*	0	2	10	20	44	49	66	69	71	79	84	90
4	0	0	2	10	34	49	68	70	78	85	94	88
5*	1	6	25	40	52	63	69	74	74	81	82	74
6	1	5	12	31	41	51	67	69	76	78	81	84
L,XIV,1		0	4	8	32	48	64	70	81	90	95	91
2		0	5	8	22	33	46	53	46	56	61	68
3*		3	15	27	43	48	62	63	74	79	81	82
4		2	5	19	21	39	54	59	68	72	79	90
5		4	20	31	43	49	58	65	71	76	81	84
6		3	4	11	31	53	64	70	80	87	90	89
M,XIV,1		5	12	20	26	37	48	51	58	66	66	77
2*		3	15	27	43	48	62	63	74	79	81	82
3		7	19	27	37	48	59	61	63	72	77	80
4		0	3	6	28	43	57	64	78	86	94	88
5		3	7	17	25	39	48	62	69	74	81	90
6		1	6	14	32	40	53	56	65	69	71	59

*Item duplicated on other form.

TABLE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age 9	10	11	12	13	14	15	16	17	18
L,AA,1	0	2	7	17	29	36	50	57	67	68
2	2	5	16	24	34	41	40	49	59	54
3	0	1	8	19	30	38	54	60	72	73
4	2	6	13	27	46	49	55	60	64	61
5	0	1	3	11	20	26	44	45	59	60
6	2	4	9	17	28	39	34	52	54	68
7	2	3	13	20	30	28	42	47	53	57
8	4	6	14	30	43	43	50	52	59	56
M,AA,1	0	0	6	16	30	40	53	61	64	70
2	2	6	9	20	31	44	47	58	54	77
3	3	6	14	22	31	29	43	48	46	48
4	3	7	18	27	34	40	52	48	56	63
5	0	0	5	11	22	28	45	51	60	72
6	9	12	22	32	37	48	53	55	64	68
7	0	0	2	9	20	27	42	50	60	60
8	3	6	13	18	31	35	38	45	51	54
L,SAI,1		1	2	8	16	17	29	32	45	46
2		9	20	26	29	32	40	39	41	48
3		4	12	16	24	25	38	39	40	41
4*		4	9	15	26	31	40	39	44	50
5		0	0	5	10	15	26	35	38	36
6*		0	2	6	10	9	25	39	41	48
M,SAI,1		6	7	17	28	26	35	37	58	55
2		1	5	7	19	20	26	37	46	43
3*		0	2	6	10	9	25	39	41	48
4*		4	9	15	26	31	40	39	44	50
5		0	1	7	13	15	34	35	39	44
6		3	8	16	19	31	30	30	42	35

*Item duplicated on other form.

TABLE 26 (Cont.)

PER CENTS PASSING ITEMS BY AGE

Item	Age 11	12	13	14	15	16	17	18
L,SAII,1	1	4	7	10	17	17	22	30
2	2	5	6	9	20	22	34	40
3	3	9	12	13	19	22	19	30
4	0	3	5	6	12	12	28	41
5	3	10	17	16	29	28	21	29
6*	7	12	13	18	22	24	26	22
M,SAII,1	0	4	7	15	26	32	33	35
2	2	6	11	20	19	27	34	46
3	0	3	6	9	14	20	28	34
4	4	10	17	14	27	27	34	31
5	2	10	17	15	26	27	38	34
6*	7	12	13	18	22	24	26	22
L,SAIII,1		0	2	1	3	8	6	13
2		5	8	8	27	24	20	30
3		0	2	2	8	16	16	20
4		4	8	4	13	10	16	21
5		1	1	2	3	5	2	8
6		2	3	4	4	6	6	9
M,SAIII,1		2	3	1	8	14	14	22
2		2	3	3	4	11	7	9
3		2	4	8	16	16	16	24
4		3	5	5	8	6	10	9
5		0	2	1	1	7	3	9
6		2	4	3	10	17	12	16

*Item duplicated on other form.

Chapter IX

FACTOR ANALYSES

It is to be regretted that facilities were not available for the use, necessarily extensive, of the factorial methods as supplementary criteria for the rejection or retention of test items. Without committing ourselves to any particular theory as to the organization of intellectual abilities, we are inclined to the position that a useful measuring scale should be highly saturated with one common factor to the exclusion of all conspicuous group factors. These conditions are necessary if the scores of individuals are to be comparable - the presence of a large group factor, or factors, permits two equal scores to be qualitatively different and two different scores to be quantitatively (with respect to the central function being measured) the same. The realization of this aim to construct a scale which measures one central function, we believe, can be attained most exactly by the factorial methods, but such methods, because of the labor involved are not feasible when one is to choose, say, 30 items from a total of 100.

It will be recalled that several criteria, separately and in combination, have been employed for the purpose of rejecting or retaining items for the present revision. Certain of these criteria tend to select items which will intercorrelate, thus assuring the presence of a common factor, although not assuring the absence of group factors. In the first place, the a priori selection of items for the original tryout, the subsequent checking of each item against the 1916 Stanford Revision, and the further requirement that the item show a satisfactory per cent passing progression with age, should all operate so as to lead to the retention of items having something in common. Then the additional criterion that the item

must show a fair degree of positive correlation with the composite point score based on all the items tends to the inclusion of items which are definitely saturated with a common factor. In fact, it has been shown[1] that to a close degree of approximation the correlation of an item with the total score corresponds to its first (or general) factor loading.

The purpose of the factor analyses herein reported was twofold: to supply an objective answer to the question as to whether the items retained (by other criteria) are saturated with a common factor to the exclusion of group factors; and possibly to supply information which might suggest whether the common factor at one age level corresponds to the common factor at another age level. Or stated differently, the Thurstone centroid method of factorization has been used in order to answer in full or in part the question: Do the items of a given level measure a common central factor, and, if so, is the central factor at one level of maturity the same as at other levels?

In view of the criteria used for the retention or rejection of items, one need not be surprised if it is found that a general factor runs through the items at a given level. If such a common factor is discovered, and further, if no group factors are found, it does not follow that the results lend support to a two-factor theory of intelligence. If, on the other hand, the analyses indicate that group factors, as well as a general factor, are present, this could be taken as evidence against the two-factor theory since the selection of the retained items has been most favorable to the finding of a single common factor. The reader, however, should understand that our purpose here is the analysis of a given set of variables (items), and that little can be generalized from our limited setup to the broader aspects of the organization of abilities.

[1]See M. W. Richardson, "Notes on the Rationale of Item Analysis," *Psychometrika*, 1936, 1, 69-76.

It should be mentioned at the outset that, in general, one cannot expect to find large factor loadings or large common factor variances for individual items because of their relatively low reliabilities. In other words, each variable or item will yield a rather large error variance which coupled with possible specific variance will tend definitely to limit its communality. It has not been possible to determine the reliability of each of the many items, but two lines of approach may suggest the possible magnitude of the reliability coefficients. Thus for some 34 items for which alternate 'forms' (one or more) are available, we have found for the several age or experimental groups a total of 112 'reliability coefficients' (tetrachoric r's). These coefficients have a median value of .65, and 80 per cent of them are between .45 and .85. We have no way of knowing whether these 34 items are representative so far as reliability is concerned.

The second indication of an item's reliability is the inference that can be made from its correlation with other items. If one fallible item correlates .70 with another fallible item, it is safe to assume that it has a reliability in the vicinity of .70 or higher. Thus the maximum correlation which an item shows with another item might be taken as a rough approximation to its reliability. A tabulation from the several tables of intercorrelations of these maximum values shows a median value of .66 for the possible 368 maximum r's with 90 per cent falling between .45 and .85. The reasons for the total number, 368, exceeding the number of tests, 258, in the two forms is that a large number of items appear twice, i.e. in two different tables of intercorrelations. This point will become clear when the general plan for the factor analyses is described.

From the foregoing it would seem safe to say that in general the item reliability is near .65, with considerable variation above and below this value. Insofar as we are concerned with the average common factor variance

for a given set of items, it can be said that it is restricted by the presence of error variance which amounts, on the average, to about 35 per cent of the total item variance, and accordingly the communalities are limited, on the average, to .65.

In planning the setup for the factor analyses, it was our aim to have each item of the scale included in at least one analysis, and to have a series of overlaps. For example, the analysis based on the 2-year-old sample, or experimental age 2, includes the items from both forms which are located at year levels II and II-6, while the analysis based on experimental age 2-1/2 includes the items at levels II-6 and III. Thus the items at level II-6 overlap or are common to these two analyses. Similarly, the items at year III are common to the analyses based on 2-1/2-year-olds and on 3-year-olds, and so on up the age scale to the analysis at experimental age 9, where the setup differs from those at the lower age groups in that it includes items at adjacent levels VIII and X. The items at level VIII are therefore common to the analyses at ages 7 and 9, while the items at level X overlap with the analyses at ages 9 and 11. This scheme of including items at both levels adjacent to the experimental age was followed for the four separate analyses based upon ages 9, 11, 13, and 15, while experimental age 18 includes the items of the three superior adult levels. The particular arrangements just described resulted from an attempt (1) to have each item included in at least one analysis with item overlap between analyses, (2) to economize as to the number of analyses necessary, and (3) to avoid extreme dichotomies in per cent passing or failing so as to have tetrachoric correlation coefficients of optimum sampling stability.

Table 27 summarizes the setups used for the fourteen separate factor analyses. The right-hand column shows the location of the items involved in a particular analysis. Also given in this table are the actual number of items used in each analysis, and the number

TABLE 27

SUMMARY OF PLAN
FOR THE SEVERAL FACTOR ANALYSES

Experimental age	N	No. of tests included	No. of overlapping tests	Location of Items
2	100	19		II, II-6
			7	
$2\frac{1}{2}$	100	19		II-6, III
			11	
3	99	25		III, III-6
			13	
$3\frac{1}{2}$	100	26		III-6, IV
			11	
4	100	26		IV, IV-6
			13	
$4\frac{1}{2}$	100	26		IV-6, V
			13	
5	100	24		V, VI
			11	
6	200	24		VI, VII
			12	
7	200	24		VII, VIII
			11	
9	200	35		VIII, IX, X
			10	
11	200	30		X, XI, XII
			9	
13	200	30		XII, XIII, XIV
			7	
15	100	30		XIV, AA., S.A.I
			8	
18	100	30		S.A.I, S.A.II, S.A.III

of items common to two analyses. It will be noted that the number of items for, say, the 2-year-group is only 19, whereas there are 28 items, including the alternates, in both forms at levels II and II-6. This and other apparent discrepancies in the table between the number of tests available and the number actually used need explanation. For this purpose it should be noted that although there are a total of 258 tests in the two forms combined, there are in fact only 189 test situations, some of which are duplicated from form to form and some of which recur with different passing standards within the same form. It should also be remarked that any two tests which bear the same name are not necessarily duplicate or recurring, since some test situations may occur twice, not exactly as the same situation but as an alternate form. Now, as regards the factor analyses, it is obvious that a test which is duplicated can be used only once for a particular analysis. Thus tests M, XIII, 5 and L, XIII, 4 are identical with identical scoring, and therefore this test situation can be used as just one variable.

In case a test at level II recurs at level II-6 with a different passing standard, it will appear only once in the analysis on experimental age 2. The reason for this should likewise be obvious — the performance on only one test situation though scored differently cannot be regarded as yielding two experimentally independent variables, and of course the correlation between the two will be spuriously high[1]. The choice as to which of two such 'tests' should be included as representing the 'test situation' was made on the basis of the one yielding the better dichotomy, i.e. nearest 50-50 for per cent passing or failing. It should, perhaps be pointed out that the exclusion of, for example, test L, III-6, 2 from the analysis on age group 3 because it recurs from level III does not

[1]This important point was not considered in the paper by R. E. Wright, "A Factor Analysis of the Original Stanford-Binet Scale," *Psychometrika*, 1939, 4, 209-220.

affect its inclusion as L, III-6, 2 in the analysis based on age 3-1/2.

All the discrepancies in Table 27 between the number of used and available tests are explicable on the basis of either duplication or recurrence with five exceptions: tests L, II-6, 3; M, II, a; M, II-6, 3 for the analysis on the 2-year-old group, and test L, VI, 1 for the analysis on age group 5. These items were not included because, due to their faulty placement in the provisional forms of the scale, they were too often omitted during the standardization-testing. Test L, II-6,a was excluded from the analysis at age 2-1/2 because of a rather extreme dichotomy which was accompanied by four-fold tables with a frequency of zero in one cell. Such tables lead to tetrachoric r's which are greater than unity and therefore of doubtful meaning.

It will also be seen from Table 27 that all N's have been reduced to 100 or 200 (for age 3 only 99 cases were available). This was done in order to facilitate the determination of the tetrachoric coefficients. The dropping of a few cases to give even N's was not exactly random in that cases with more incomplete records were dropped first. Some exceptions to these N's should be noted. At experimental age 6, tests L, VI, 1 and M, VII, 2 were included with N's of 183 and 167 respectively, and for age group 7, tests L, VII, 3; M, VII, 1; and M, VII, 2 were included with reduced N's of 190, 184, and 182. These smaller N's resulted from the fact that these items were not administered to all the subjects at these ages because of faulty placement in the provisional forms. It seemed reasonable to include these items since the N's were still fairly large, whereas for the items mentioned in the previous paragraph the reduced N's were too small to justify their inclusion.

Although certain tests were not included in certain analyses, we have succeeded in having each test situation included in at least one of the fourteen analyses, and at the same time have avoided extreme dichotomies. As a

matter of fact, 93 per cent of the dichotomies are between 20 and 80 per cent for passing (or failing), one-half are between 35 and 65, and only four are more extreme than 10-90. Two of these four are 9-91, the other two, 8-92.

Perhaps a word should be said concerning the precautions taken to guarantee the accuracy of the computation involved in determining 4780 tetrachoric coefficients and in extracting three centroid factor loadings for fourteen tables of intercorrelations based on from 19 to 35 variables. Hollerith cards, 1900 in number, were punched by a trained operator and at a later time a duplicate set was punched by the same operator; then the two cards for an individual were checked one against the other. The agreement of the several marginal totals was an absolute check on the fourfold table frequencies as copied from the Hollerith sorter, and the internal checks provided by the centroid method are sufficient to insure computational accuracy for the factor loadings.

According to the criterion first proposed by Thurstone, one should continue to extract factors until the residual variance is equal to or less than the sampling variance of the original intercorrelations. This criterion was used by us at the time (1937) these analyses were carried out, with the result that one factor seemed sufficient for each of the 14 separate analyses. Since the adequacy of the residual criterion had been questioned, we proceeded to extract two more factors. It was thought at that time that the extraction of second and third loadings might be superfluous, a mere dallying with chance, but it was done lest a more exact criterion be devised which would call for at least three factors. The writer has since found a more reasonable criterion[1] which is based on an adjustment to the residuals in order to make them more analogous to partial correlations. This criterion required that the variance of the residuals, as

[1]See *Psychometrika*, March, 1942.

partials, shall approach the sampling variance of a zero correlation based on the given N.

The application of either the ordinary, or the partial, residual criterion to data based upon tetrachoric correlations is greatly complicated by the fact that the sampling variance of tetrachorics is partly a function of the dichotomies. Obviously there will be no single critical value to be approached by the variance of the residuals. We have therefore considered two values as representing possible magnitudes for the sampling variance, one based upon typical cuts of 35-65, and one based upon the more extreme 15-85 cuts. Now, some few r's will have errors somewhat smaller (as much as .009) than obtain for 35-65 cuts, while other r's will have errors as much as .05 larger than hold for 15-85 cuts.

The complete data concerning the standard errors of zero tetrachorics, the standard deviations of the ordinary residuals and of the residuals as partials are presented in Table 28. An examination of this table certainly emphasizes the need for a valid criterion for the number of factors, especially when matrices of tetrachoric correlations with varying dichotomies are being analyzed. Consideration of either the ordinary or the partial residual S.D's makes one rather dubious as to whether the extraction of more than one factor can be justified. Perhaps a second factor is needed for the analyses at ages 2, 2-1/2, 6, 7, 11, and 18; but it seems unlikely that a third factor would contain anything more than chance. We are, nevertheless, presenting loadings for three factors for each analysis.

There is a point of considerable interest which results from a perusal of Table 28. It will be noted that the first, second, and third factor residual distributions consistently show less variation for those age groups where N equals 200 than where N equals 100. This might have been surmised - the reduction of sampling errors has definitely reduced the residuals and therefore the magnitudes of successive subsequent factor loadings. A

TABLE 28

STANDARD DEVIATIONS OF ORDINARY AND PARTIAL RESIDUALS COMPARED WITH ERRORS DUE TO SAMPLING

Age group	Standard error for zero order tetrachorics		Standard deviations of residual distributions					
			Ordinary			As partials		
	35-65	15-85	1st	2nd	3rd	1st	2nd	3rd
2	.166	.235	.148	.121	.107	.270	.267	.272
2	"	"	.150	.110	.104	.240	.213	.232
3	"	"	.134	.119	.110	.215	.217	.223
3	"	"	.134	.119	.103	.232	.237	.235
4	"	"	.128	.117	.102	.217	.222	.220
4	"	"	.142	.128	.115	.224	.228	.234
5	"	"	.146	.129	.119	.230	.232	.239
6	.117	.166	.113	.099	.090	.174	.169	.167
7	"	"	.102	.093	.078	.181	.183	.172
9	"	"	.108	.095	.090	.166	.160	.162
11	"	"	.106	.088	.077	.171	.157	.151
13	"	"	.106	.096	.085	.172	.171	.166
15	.166	.235	.120	.105	.097	.239	.241	.249
18	"	"	.166	.132	.120	.264	.253	.259

second factor loading, for example, of .40 is not as real at age 5 (N, 100) as at age 6 (N, 200). As will be seen later, the per cent contribution to test variance for the second and third factors is not as great when N equals 200 as when N equals 100, whereas the variance contribution of the first factor seems to be independent of N. On a priori grounds it does not seem reasonable to believe that the items involved in the analyses at ages 6, 7, 9, 11, and 13 (N's of 200) are such as to show a factorial structure consistently different from that for the other analyses. Certainly the items in the analysis on age group 5 are markedly similar to those in the analysis at age 6 (11 items are in common, and 6 highly similar item situations are also in common), and the analyses at ages 13 and 15 are likewise based upon items which are much alike; yet the residual variations for ages 5 and 6 and for ages 13 and 15 differ in the direction of greater variation for the smaller samples. The percentage contribution of the second factor to test variance also shifts as we pass from age group 5 to 6; when N is smaller, the second and third factors seem to contribute more to test variance. All these facts tend to suggest that the second and third factor loadings are not as stable from the sampling standpoint as the first factor loadings, and that the general magnitude of residuals and subsequent loadings is a function of the size of the sample.

It must be suspected that attempts to rationalize the meaning of the third factor loadings, and to a large extent the second factor loadings, have in general led to grief, and consequently to the conclusion that the criteria used are not too crude. This general problem regarding the influence of sampling errors will again confront us as we proceed with an exposition of the further findings of the several analyses.

The results, in terms of centroid factor loadings, for the several analyses, are presented in Tables 29-42. The name of the test or item is preceded by its location

as to form and age level and followed by its three factor loadings. The right-hand, apparently incomplete, column gives the first factor loadings for the overlapping items as computed on the next higher experimental age group. Thus the column in Table 29 headed as k_1 (2-1/2) gives the k_1 values found in the analysis at age 2-1/2 for those items which are common in the 2- and 2-1/2-year analyses. At the bottom of each table will be found $\Sigma k^2/n$, or the average proportion of test variance explicable on the basis of each of the centroid factors.

In considering the results, let us first turn our attention to the question as to whether, for a given level, the items are measuring a general common factor to the exclusion of group factors. An examination of the first factor loadings in the several tables shows that, for a given analysis, all the items are saturated with a common factor. The importance of these common (first) factors, one for each analysis or fourteen in number, varies somewhat from analysis to analysis, but in general they account for about 40 per cent of the variance of the tests. When it is recalled that on the average about 35 per cent of the test variance is due to unreliability, the significance of these common (first) factors becomes more apparent and real.

It seems safe to say that the items at a particular maturity level do measure one common function, but study of the first factor weights reveals the fact that some items are definitely low in general factor saturation. For example, item M, IV, 2, 'stringing beads,' which has a loading of only .241 for the analysis at age 3-1/2, cannot be considered a very satisfactory item so far as the measurement of a unitary trait is concerned. Neither can 'block counting,' M, X, 1 (see Tables 38 and 39) be adjudged as a suitable item. For any one given analysis there is considerable variation from item to item in first factor loadings. How much of this is real and how much is due to the sampling errors in the original experimental correlations cannot be definitely stated. The

writer, however, has published[1] empirical data on the sampling errors of centroid factor loadings which indicate that a first factor weight is subject to sampling fluctuations of about the same order of magnitude as a correlation coefficient of the same size, and of the same type, product moment or tetrachoric, as in the starting matrix.

It may be of some interest to list the items or item situations which tend to yield high first factor loadings and those which yield low ones. In the case of items or item situations for which two or more loadings are available (i.e. overlapping tests, recurring tests, and recurring test situations), it was required that the loadings be consistently high, or low, in order to be included in the listings. 'High' and 'low' are to be taken in a relative sense. Rather than throw items from all levels together, the listings are made separately for II to IV-6, V to XI, and XII to S.A.III. Each of these trichotomies includes about one third of the items in the scales.

At levels II to IV-6, high first factor loadings were yielded by these items:

Picture vocabulary
Identifying objects by name
Response to pictures
Comparison: balls; sticks
Comprehension
Opposite analogies
Pictorial identification
Materials

Low loadings occurred for the following items of levels II to IV-6:

Block building: tower
Block building: bridge
Three-hole form board: rotated
Motor coordination
Copying a circle

[1] See *Psychometrika*, 1941, 6, 141-152.

Drawing a cross
Three commissions
Stringing beads

The items or item situations at levels V to XI which are highly saturated with the general factor are:

Pictorial likenesses and differences
Similarities: two things
Vocabulary
Verbal absurdities
Similarities and differences
Naming the days of the week
Dissected sentences
Abstract words

Those at levels V to XI with low loadings are as follows:

Paper folding: triangle
Patience: rectangles
Copying a bead chain
Copying a bead chain from memory
Picture absurdities
Word naming
Word naming: animals
Block counting

Among the items at the upper levels, XII to S.A. III, the following tend to have the highest first factor loadings:

Vocabulary
Verbal absurdities
Abstract words
Differences between abstract words
Arithmetical reasoning
Proverbs
Essential differences
Sentence building

The items at the upper levels which are least saturated with the central factor are:

Problems of fact
Copying a bead chain from memory

Memory for stories
Enclosed box problem
Papercutting
Plan of search
Repeating digits
Repeating digits: reversed

When we turn to the question of the presence of group factors we are again baffled by ignorance concerning the possible sampling variation in factor loadings. The empirical study referred to above shows that second and third centroid factor loadings have standard errors which are much greater than those of correlations of the same magnitude. Two samples, each of size 100, may lead to second factor loadings for a given variable which differ by as much as .50 or .60. These considerations, coupled with the possible insignificance of the second and third factors as discussed in relation to the criteria for determining the number of factors to be extracted, force one to be skeptical as to what can be said concerning group factors in this study. Nevertheless, whenever the descriptions of items in terms of their second and third factor loadings show logical consistencies, the fact cannot be ignored, but of course it must be remembered that chance can so operate as to lead to apparent consistencies. Before discussing the possible meaning of the second and third factors for the separate analyses, it should be noted that on the average the second factor accounts for from 5 to 11 per cent of the test variance and the third contributes from 4 to 7 per cent.

When the items in the analysis at age 2 are plotted with reference to the second and third centroid axes, one gets a vague impression that the second factor differentiates slightly between items which involve some kind of 'identifying' or 'knowing' of objects and items of a 'motor' or 'memory' nature, while these latter types are roughly separated along the third axis. If we let the second axis be the abscissa and the third the ordinate (a scheme followed in subsequent discussion), the tests falling in the

two right-hand quadrants involve 'identifying,' while the upper left quadrant contains 'motor' tests and the lower left contains the two digit tests. A plot of the tests included in the analysis at age 2-1/2 shows this same general feature provided the arbitrary centroid axes are rotated counter-clockwise through about 90 degrees. Here again, however, the 'motor' tests seem to merge with the 'identifying' tests. That these results may not be entirely due to chance or sampling is supported by the fact that the variance of the first factor residuals (see Table 28) indicates that one factor may not be sufficient to explain the intercorrelations on age groups 2 and 2-1/2. It should also be noted that the second factor at age 2 accounts for an average of 9.5 per cent of the test variance and that 10.9 per cent at age 2-1/2 is accounted for by the second factor. (It is not assumed or implied that the second factor as found on one age group is the same as that found at another age.) For no other analysis except at age 18 does the second factor contribute so much to the test variance.

The tests in the analyses at ages 3, 3-1/2, 4, and 4-1/2 when plotted with reference to the respective second and third centroid coordinates fail to fall in any logical groups or form clusters or show simple structure, and therefore no meaning can be attached to the second and third centroid axes or rotation thereof. If group factors exist among the items at these levels of maturity, our samples are too small to demonstrate them.

A plot of the second and third loadings for the items in the analysis at age 5 reveals some order: the upper right quadrant contains items which are, relative to the other items, more 'verbal' in nature (two pictorial tests in this same quadrant nullify somewhat the meaning of this cluster), and the upper left quadrant contains four 'numerical,' number concept, tests. It seems impossible, however, to make any sense out of the arrangement of the tests in the lower quadrants. This differentiation between 'numerical' and the more 'verbal' tests

stands out rather clearly in the analysis based on the 6-year-old group, where the separation is again along the second (arbitrary) axis. The third axis, as at age 5, seems void of meaning. It will be recalled that the spread of the first factor residuals (see Table 28) for the analysis on age group 6 was such as to suggest the extraction of a second factor.

There is a lack of consistent groupings for the tests in the analysis on age 7, while for age 9 the 'verbal' type tests seem opposed, along the second axis, to the 'memory' type tests. This same separation occurs for the 11-year-group, and therefore this differentiation of the items located at age levels 8, 9, 10, 11, and 12 is not likely to be the result of chance. For both analyses the 'repeating of digits' tests are roughly opposed along the third axis to tests like 'memory for designs,' but in general it is difficult to assign any meaning to the third centroid factor.

When we turn to the results for age 13, we find an inconspicuous tendency for the 'verbal,' the 'problem' and the 'memory' types of tests to be separated, but the scattering and merging are such as to preclude any very definite conclusions concerning a group factor. A similarly vague differentiation between 'problem' and 'verbal' tests is found for the items in the analysis on experimental age 15.

The factorial results based on age group 18 for the analysis of the items at the three superior adult levels are the most disconcerting so far as the possible presence of a sizable group factor is concerned. The residual variance (see Table 28) indicates that more than one factor is needed to explain the intercorrelations, and, as is seen from Table 42, the second centroid factor contributes an average of 10.7 per cent of the test variance. This second factor appears to involve the difference between 'verbal' items and tests of immediate memory such as repeating digits. In fact, the four tests which require the repeating of 8 or 9 digits are more

highly saturated with the second than with the first or more general factor. Even though it is fairly easy to pick out this conspicuous, though small, digit cluster, it is difficult if not impossible to assign definite meaning to the haphazard pattern formed by the remaining items in the analysis on the 18-year-group.

The foregoing consideration of the outcome of the several factor analyses leads to the definite conclusion that all the items in a particular analysis are saturated in varying degrees with a general factor, and to the tentative conclusion that one factor is sufficient to account for the intercorrelations except those based on experimental ages 2, 2-1/2, 6, 18, and possibly 7 and 11. The items at each of these levels apparently involve one or more group factors, but in no case is the evidence sufficiently clear-cut to justify any elaborate deductions regarding the nature of these possible group factors. We have already suggested provisional characterization for these item groupings — to say more would require analyses based upon much larger samples.

It should also be remarked, perhaps at this place, that in a given analysis any two or more items which represent alternate 'forms' of the same 'test situation' tend to hang together so far as their factorial description is concerned, and thus every such set of two or three items will form a cluster, and, conceivably, it might be said that this indicates the presence of a group factor since more than one item is involved. The writer is not willing to speculate as to whether additional centroid axes or rotations thereof would eventually lead to the isolation and elevation of such a small cluster to the status of a psychologically meaningful factor. One can be sure, however, that these small 'group factors' could not contribute sufficiently to I.Q. variance to invalidate the comparability of I.Q.'s of the same magnitude for individuals of approximately the same life-age. Unfortunately, this latter statement cannot be made with regard to the 'group factors' which seem to emerge at age 2 and

2-1/2, at age 5 and 6, and at age 18. These 'group factors,' the existence and importance of which may be questioned, appear to be of sufficient prominence to cause small, though not inconsequential, qualitative differences between the I.Q.'s of two individuals when the mental maturity of either or both is at any one of the levels where these factors emerge.

The absence of conspicuous group factors, though necessary, is not a sufficient condition either for the generalization that the I.Q.'s for individuals of differing maturity levels are strictly comparable or for generalizations regarding the relative constancy of I.Q.'s. In either case we must have the added condition that the same central function is being measured by the items at the various maturity levels. Certain of the results of the factor analyses which we have already discussed tend to show that each item at a given level measures a factor or function which is common to all the items at that level, and we will now consider evidence, not conclusive though certainly more than presumptive, which indicates that the common or first factor at one level is the same as at other levels.

The first approach to this problem is by way of the series of overlapping tests, i.e. the tests which are common to any two adjacent analyses. The general factor as found for the analysis on age group 2 is the factor common to the items located at age levels II and II-6, while the general factor as found for experimental age 2-1/2 is common to the items at levels II-6 and III. The items at age level II-6, which overlap or are included in these two analyses, will have two sets of factor loadings, one for the age 2 analysis and the other for the age 2-1/2 analysis. In attempting to answer the question as to whether or not the common factor at one age represents the same function as that found at the next higher age we are faced with four alternatives: factors the same or different and the two sets of loadings in agreement or disagreement. Let us examine these four possibilities.

If the factors differ, it seems logical to assume that the two sets of loadings will also differ; and if the factors are the same, it seems reasonable to assume that the loadings will agree. But it is the converse of these propositions which must concern us since the available data yield information concerning the concurrence of the two sets of first factor loadings for the overlapping tests. If the loadings show disparities which are greater than are to be expected on the basis of sampling, it would seem safe to suppose that the two factors are not identical. If the two sets of loadings correspond, within limits of the sampling errors, can it be said that the two common factors are homologous? This, it seems to the writer, can be answered in the affirmative by way of a negative: if the two factors really differed, the loadings would not agree.

If the above logic is sound, we can proceed to a study of the behavior of the first factor loadings for the several series of overlapping tests. Before doing this, it should be recalled that the standard error of a first centroid loading is, according to our empirical results, about the same as that for a correlation coefficient of the same magnitude. A value of .60 may be taken as representative of the first factor loadings. A tetrachoric r of this magnitude based on typical 35-65 cuts would have a standard error of .121 for N = 100, and .085 for N = 200. Presumably loadings higher than .60 would have smaller sampling errors, while lower loadings would have larger errors. To be on the conservative side for such conclusions as we draw, let us take smaller values, .10 and .07 for N's of 100 and 200 respectively, as approximations for the sampling errors of the first factor loadings in this study. Thus the standard error of the difference between two loadings based on two different samples with N's of 100 each will be about .14, with N's of 200 each about .10, and when one sample contains 100, the other 200, about .12.

We are now ready again to peruse Tables 29-41,

this time to compare the two sets of loadings for the tests which overlap two analyses, i.e. the k_1 values in the right-hand column with the first column of figures. In general the loadings for all 13 groups of overlapping tests show marked agreement. Of the 136 differences, 91 are less than .10, and 120 are less than .20. There are, however, a few rather large, though not necessarily statistically significant, discrepancies which should receive attention. We will discuss these in the order in which they occur. In Table 29 tests L, II-6, 2, 'identifying parts of body' and M, II-6, a, 'stringing beads,' and in Table 30 test L, III, 3, 'block building: bridge,' yield differences which are about 1.5 times their standard errors. These differences may be real, but it is difficult to explain why presumably similar tests do not also show differences.

For the tests which overlap the analyses at ages 3 and 3-1/2 (see Table 31), there are three differences (for items L, III-6, a; L, III-6, 1; and L, III-6, 5) which are possibly non-chance, and one difference (item M, II-6, a) which is definitely significant. Although the writer sees no a priori reason for these discrepancies, it cannot be argued that they are not real. But the fact that the alternate 'form' (M, III, 3) of test L, III-6, 5 yields a loading of .641 (age 3 analysis) compared with .419 for L, III-6, 5 makes one skeptical as to the reality of a .27-point difference. The next possible non-chance difference occurs for test M, IV-6, 2, 'definitions,' with loadings of .649 (age 4 analysis) and .431 (age 4-1/2 analysis), as can be seen in Table 33. An alternate 'form' of this test occurs as L, V, 3 in the analyses at ages 4-1/2 and 5 with loadings of .427 and .703 respectively. If these were accompanied by shifts in other item loadings, it might be suggested that the common factors running through the items in the analysis at age 4-1/2 may differ somewhat from the factors found in the adjacent analyses. A difference which is about twice its sigma will be noted in Table 36 for test L, VII, 2,

'similarities: 2 things.' The fact that the other thirteen differences in this table are small and insignificant tends to overshadow this one difference. A statistically significant difference will be found in Table 37 for test M, VIII, 6, 'opposite analogies,' but here again the remaining differences are so small, averaging less than .05, that it seems illogical to suppose that the two common factors are different. The analyses at ages 9 and 11 (see Table 38) give one rather large, and likely non-chance, discrepancy – test M, X, 2, 'memory for stories.'

Thus out of a total of 136 pairs of loadings for overlapping tests, only 12 show differences large enough to attract attention, and of these 12, only 3 seem to possess statistical significance as judged by approximate, and very likely underestimates of the sampling errors. That so close an accord is found for two sets of factor loadings for overlapping variables is indeed surprising, especially when it is remembered that we have not only moved tests from one 'battery' or setup to another but have also made the analyses on different samples. These findings certainly lead to the belief that the general factors at two adjacent levels are identical or nearly so. As regards the centroid factors found for non-adjacent experimental ages, it would seem likely that they too are homologous, but one cannot here apply without reservations the mathematical principle that things equal to the same thing are equal to each other. Nevertheless, it seems to the writer that the findings for the overlapping tests are so highly suggestive as to lead to the conclusion that the common factors for the several levels are nearly equivalent.

A second line of reasoning will now be considered which tends in general to support the above conclusion and at the same time to cast some doubt on its validity. It will be recalled that certain tests recur with different passing standards and that other tests recur as alternate 'forms' of the same situation. A recurring test is included only once, while a test may have one or more

alternate 'forms,' within an analysis. Such tests or test situations may appear at several different levels, and thus we are provided with additional overlaps which are not confined to adjacent analyses. For example, the two 'repeating digits' items which are included in the analysis on age 2 overlap as identical tests in the analysis on age 2-1/2, which, in turn, includes two other 'repeating digits' tests, etc. Let us examine the behavior of the first factor loadings for such of these tests as occur (recur) in four or more analyses. Whenever within a given analysis a test situation occurs more than once we shall average its first factor loadings in order to have a more typical weight.

The results of this approach to the problem of the equivalence of the general factors being measured at the various levels will be found in Table 43, in which an omission means that the test situation did not occur in a particular analysis. The general tenor of this table is the marked consistency in loadings for a given test as we follow it from lower to higher levels. It seems to the writer that such accord is not due to chance or accident, and that support is herewith found for the conclusion that the several common factors are homologous. There are, however, a few trends in Table 43 which force us to modify this conclusion. It will be noted that 'repeating digits' and 'picture vocabulary' have higher loadings at ages 2, 2-1/2, and 3 than at later ages, that 'opposite analogies' has somewhat higher values at 3-1/2, 4, and 4-1/2 than later, and that the loadings for 'vocabulary' gradually rise from age 6 to 18. These changes may indicate a change in the nature either of item situations or of the general factor as we pass from lower to higher levels, or both. For instance, 'repeating digits' may at the early ages depend upon whatever ability is involved in attention and in following directions, whereas at later ages it may depend more upon immediate memory. As regards a possible change in the general factor, it may be argued that the verbal element becomes

more conspicuous at the higher levels and that the behavior of the loadings for 'vocabulary' supports this possibility. It seems quite logical to the writer to believe that some differences do exist in the common factor called for at various age levels. This is a tentative conclusion, and one which applies to changes in the nature of the test items rather than to changes in individuals as one passes from childhood to maturity.

The results of the several analyses reported in this chapter may be conveniently summarized under three headings: (1) The several items at a particular maturity level are saturated, though in varying degrees, with a factor which is general or common to the tests located at the given level. When one considers the unreliability of single items the amount of variance due to the several first factors, one for each analysis, is reasonably satisfactory. (2) The first factor for each analysis is sufficient to account for the intercorrelations except in the analyses based on experimental ages 2, 2-1/2, 6, and 18. At these levels there seems to be some evidence that certain items are measuring more than a single factor plus specifics, but our limited samples are insufficient for definitely establishing the importance or nature of the possible group factors at these levels. (3) The first factor loadings for tests which overlap two adjacent analyses, and for tests and test situations which recur or run though several analyses, show a high degree of conformity. This, despite some conflicting data, has been interpreted as showing that the common factors found at the several levels are nearly identical. (A fourth, but not discussed, finding should also be mentioned: there are no observable differences between the items in Forms L and M as to their factorial structure.)

The implications of these findings need to be carefully qualified. It would appear that we have evidence for one of the conditions (absence of group factors) necessary for the quantitative and qualitative comparability of I.Q.'s for individuals of about the same mental ma-

turity level, but the testing procedure is such that an individual's failures and successes may spread over a greater range of tests than is included in our separate analyses. Thus if group factors exist within this larger range of items, small qualitative differences in I.Q.'s will be possible. This point could be investigated, not by factorial methods, since larger ranges of items than we have used would involve extreme dichotomies of passing or failing, but by an analysis of the patterns of items failed or passed by those individuals whose spread of performance is greatest (very large samples would be needed for such an analysis). A second implication, based on the conclusion that there is just one common factor at each level and also that the several general factors found at the various levels are identical or nearly so, is that the I.Q.'s for individuals of differing mental-maturity levels or for the same individual at different stages of development are comparable quantitatively and qualitatively. This should not be construed as an argument for I.Q. constancy, but rather as evidence that a condition necessary for I.Q. constancy has been met. If it is found by a follow-up study that the I.Q.'s of six-year-olds correlate highly with their I.Q.'s as of age 2, the finding would support our conclusion concerning homologous common factors. If, however, such a correlation were fairly low, our conclusion might be open to suspicion but not proved invalid, since other conditions, such as inequalities in maturation rate or differences in experience or state of health, etc., may be responsible for the apparent lack of constancy.

TABLE 29

FACTOR LOADINGS FOR ANALYSIS AT AGE 2

Location	Name of test	k_1	k_2	k_3	$k_1(2\frac{1}{2})$
L,II,a	Obey. simple commands	.580	.408	.191	
L,II,1	3-hole form board	.715	-.004	.137	
L,II,4	Block building: tower	.444	-.502	.240	
L,II,5	Picture vocabulary	.764	.360	-.484	
L,II,6	Word combination	.707	-.240	-.168	
L,II-6,a	Identify. obj. by name	.765	.227	-.192	-
L,II-6,1	Identify. obj. by use	.680	.261	.200	.657
L,II-6,2	Identify. parts of body	.713	-.112	-.240	.455
L,II-6,5	Repeating 2 digits	.744	-.294	-.369	.695
L,II-6,6	3-hole fm. bd. rotated	.460	-.411	.327	-
M,II,1	Delayed response	.497	.090	.095	
M,II,2	Identify. obj. by name	.809	.518	.153	
M,II,3	Identify. parts of body	.800	.225	-.233	
M,II,5	Picture vocabulary	.763	.375	-.222	
M,II-6,a	Stringing beads	.655	-.249	-.029	.438
M,II-6,1	Identify. obj. by use	.780	-.030	.201	-
M,II-6,2	Motor coordination	.378	-.104	.226	.453
M,II-6,5	Repeating 2 digits	.645	-.523	-.200	.677
M,II-6,6	Obey. simple commands	.638	.028	.370	.592
	$\Sigma k^2/n$	.451	.095	.061	

TABLE 30

FACTOR LOADINGS FOR ANALYSIS AT AGE $2\frac{1}{2}$

Location	Name of test	k_1	k_2	k_3	$k_1(3)$
L,II-6,1	Identify. obj. by use	.657	.456	-.265	
L,II-6,2	Identify. parts of body	.455	-.257	-.299	
L,II-6,3	Naming objects	.747	-.339	-.423	
L,II-6,5	Repeating 2 digits	.695	-.506	.352	
L,III,a	3-hole fm. bd. rotated	.342	.294	-.248	.513
L,III,2	Picture vocabulary	.761	.026	-.366	.816
L,III,3	Block building: bridge	.313	.388	-.039	.566
L,III,4	Picture memories	.560	-.196	-.134	.495
L,III,5	Copying circle	.501	.201	.371	.332
L,III,6	Repeating 3 digits	.687	-.409	.311	.746
M,II-6,a	Stringing beads	.438	.244	-.174	
M,II-6,2	Motor coordination	.453	.007	.313	
M,II-6,5	Repeating 2 digits	.677	-.533	.128	
M,II-6,6	Obey. simple commands	.592	.179	.111	
M,III,2	Picture vocabulary	.755	-.134	-.159	.722
M,III,3	Identify. obj. by use	.656	.420	-.051	.641
M,III,4	Drawing vertical line	.458	.415	.340	.557
M,III,5	Naming objects	.775	.178	-.124	.663
M,III,6	Repeating 3 digits	.755	-.410	.226	.661
	$\Sigma k^2/n$	.375	.109	.067	

TABLE 31

FACTOR LOADINGS FOR ANALYSIS AT AGE 3

Location	Name of test	k_1	k_2	k_3	$k_1(3\frac{1}{2})$
L,III,a	3-hole fm. bd. rotated	.513	-.199	-.124	
L,III,1	Stringing beads	.476	-.119	.156	
L,III,2	Picture vocabulary	.816	.234	.194	
L,III,3	Block building: bridge	.566	.244	-.153	
L,III,4	Picture memories	.495	-.263	.226	
L,III,5	Copying circle	.332	.138	.235	
L,III,6	Repeating 3 digits	.746	-.011	-.327	
L,III-6,a	Drawing cross	.636	-.467	.337	.379
L,III-6,1	Obey. simple commands	.584	.118	-.159	.778
L,III-6,3	Comparison of sticks	.664	-.257	-.129	.760
L,III-6,4	Response to pict. I	.732	.129	-.309	.593
L,III-6,5	Identify. obj. by use	.419	.387	.321	.694
L,III-6,6	Comprehension I	.628	-.043	-.331	.663
M,III,2	Picture vocabulary	.722	.254	.149	
M,III,3	Identify. obj. by use	.641	.337	-.177	
M,III,4	Drawing vertical line	.557	-.700	-.264	
M,III,5	Naming objects	.663	.298	.087	
M,III,6	Repeating 3 digits	.661	.260	-.309	
M,III,6-a	Matching objects	.279	-.416	.337	.728
M,III-6,1	Comparison of balls	.640	-.264	.198	.755
M,III-6,2	Patience: pictures	.653	-.134	-.127	.543
M,III-6,3	Discrim. animal pict.	.603	.124	.103	.667
M,III-6,4	Response to pict. I	.773	.183	.141	.723
M,III-6,5	Sorting buttons	.590	.159	.174	.693
M,III-6,6	Comprehension I	.637	.043	-.324	.677
	$\Sigma k^2/n$	.377	.075	.054	

TABLE 32

FACTOR LOADINGS FOR ANALYSIS AT AGE 3½

Location	Name of test	k_1	k_2	k_3	$k_1(4)$
L,III-6,a	Drawing cross	.379	.296	.359	
L,III-6,1	Obey. simple commands	.778	.136	-.175	
L,III-6,2	Picture vocabulary	.632	.125	-.229	
L,III-6,3	Comparison of sticks	.760	-.230	.306	
L,III-6,4	Response to pict. I	.593	-.018	.300	
L,III-6,5	Identify. obj. by use	.694	-.152	-.174	
L,III-6,6	Comprehension I	.663	-.282	-.303	
L,IV,a	Memory for sentences I	.658	-.343	.235	.603
L,IV,2	Naming obj. from memory	.483	.334	-.294	.653
L,IV,3	Pict. completion: man	.545	-.219	.117	.503
L,IV,4	Pictorial identification	.635	-.289	-.227	-
L,IV,5	Discrimination of forms	.550	.318	-.136	.679
L,IV,6	Comprehension II	.813	-.337	-.097	.802
M,III-6,a	Matching objects	.728	.348	-.009	
M,III-6,1	Comparison of balls	.755	.165	-.035	
M,III-6,2	Patience: pictures	.543	.497	-.341	
M,III-6,3	Discrim. animal pict.	.667	.286	-.061	
M,III-6,4	Response to pict. I	.723	.161	.292	
M,III-6,5	Sorting buttons	.693	.308	-.088	
M,III-6,6	Comprehension I	.677	-.270	-.280	
M,IV,1	Picture vocabulary	.740	-.270	-.351	.650
M,IV,2	Stringing beads	.241	.260	.156	.400
M,IV,3	Opposite analogies I	.705	-.347	.401	.754
M,IV,4	Pictorial identification	.644	-.308	-.245	.546
M,IV,5	Number concept of two	.617	.119	.417	.715
M,IV,6	Memory for sentences I	.668	-.272	.221	.727
	$\Sigma k^2/n$	.422	.076	.063	

TABLE 33

FACTOR LOADINGS FOR ANALYSIS AT AGE 4

Location	Name of test	k_1	k_2	k_3	$k_1(4\frac{1}{2})$
L,IV,a	Memory for sentences I	.603	-.249	.212	
L,IV,1	Picture vocabulary	.701	.145	.206	
L,IV,2	Naming obj. from memory	.653	.291	.035	
L,IV,3	Pict. completion:man	.503	.426	.235	
L,IV,5	Discrimination of forms	.679	.154	-.090	
L,IV,6	Comprehension II	.802	-.192	-.226	
L,IV-6,a	Pictorial identification	.728	.229	-.188	.718
L,IV-6,1	Aesthetic comparison	.602	.137	-.173	.623
L,IV-6,2	Repeating 4 digits	.435	-.291	.419	.533
L,IV-6,3	Pictorial like. and diff.	.470	-.211	.124	.578
L,IV-6,4	Materials	.660	.281	-.281	.692
L,IV-6,5	Three commissions	.445	-.175	-.201	.409
L,IV-6,6	Opposite analogies I	.778	-.190	-.207	.799
M,IV,1	Picture vocabulary	.650	.079	.137	
M,IV,2	Stringing beads	.400	.201	-.650	
M,IV,3	Opposite analogies I	.754	-.076	-.245	
M,IV,4	Pictorial identification	.546	.084	.203	
M,IV,5	Number concept of two	.715	.315	-.226	
M,IV,6	Memory for sentences I	.727	.089	.330	
M,IV-6,a	Patience: pictures	.565	-.363	.291	.547
M,IV-6,1	Discrim. animal pict.	.501	.273	.188	.440
M,IV-6,2	Definitions	.649	-.350	-.109	.431
M,IV-6,3	Repeating 4 digits	.697	-.424	.300	.740
M,IV-6,4	Picture completion: bird	.611	.282	.266	.646
M,IV-6,5	Materials	.774	-.208	-.199	.719
M,IV-6,6	Comprehension II	.757	-.213	-.193	-
	$\Sigma k^2/n$	.411	.061	.065	

TABLE 34

FACTOR LOADINGS FOR ANALYSIS AT AGE 4½

Location	Name of test	k_1	k_2	k_3	$k_1(5)$
L,IV-6,a	Pictorial identification	.718	.117	.173	
L,IV-6,1	Aesthetic comparison	.623	-.213	-.158	
L,IV-6,2	Repeating 4 digits	.533	-.148	-.307	
L,IV-6,3	Pictorial like and diff.	.578	-.051	.441	
L,IV-6,4	Materials	.692	-.107	.230	
L,IV-6,5	Three commissions	.409	-.384	-.499	
L,IV-6,6	Opposite analogies I	.799	.070	.415	
L,V,a	Knot	.416	.170	.320	.547
L,V,1	Picture completion:man	.633	.542	-.115	.431
L,V,2	Paper folding:triangle	.345	.415	-.361	.536
L,V,3	Definitions	.427	-.286	.411	.703
L,V,4	Copying square	.586	.147	-.384	.509
L,V,5	Memory for sentences II	.572	.038	.034	.569
L,V,6	Counting 4 objects	.723	.053	-.015	.651
M,IV-6,a	Patience: pictures	.547	-.142	-.129	
M,IV-6,1	Discrim. animal pict.	.440	.068	-.174	
M,IV-6,2	Definitions	.431	-.601	.284	
M,IV-6,3	Repeating 4 digits	.740	-.451	-.349	
M,IV-6,4	Picture completion:bird	.646	.146	-.163	
M,IV-6,5	Materials	.719	.300	.279	
M,V,1	Picture vocabulary	.657	.013	-.104	.670
M,V,2	Number concept of three	.626	.384	.140	.480
M,V,3	Pictorial sim. and diff.	.685	-.054	-.043	.610
M,V,4	Patience:rectangles	.310	-.200	-.129	.470
M,V,5	Comprehension II	.707	-.109	.203	.630
M,V,6	Mutilated pictures	.773	.331	-.096	.613
	$\Sigma k^2/n$	.366	.072	.071	

TABLE 35

FACTOR LOADINGS FOR ANALYSIS AT AGE 5

Location	Name of test	k_1	k_2	k_3	$k_1(6)$
L,V,a	Knot	.547	.109	-.110	
L,V,1	Picture completion:man	.431	.411	-.079	
L,V,2	Paper folding:triangle	.536	-.143	-.193	
L,V,3	Definitions	.703	.468	.296	
L,V,4	Copying square	.509	.026	-.591	
L,V,5	Memory for sentences II	.569	.283	.228	
L,V,6	Counting 4 objects	.651	-.254	.173	
L,VI,2	Copying bead chain mem. I	.709	-.311	-.331	.652
L,VI,3	Mutilated pictures	.564	-.154	-.190	.406
L,VI,4	Number concepts	.667	-.508	.279	.770
L,VI,5	Pictorial like. and diff.	.610	.219	.219	.726
L,VI,6	Maze tracing	.482	.067	-.178	.573
M,V,1	Picture vocabulary	.670	-.067	.077	
M,V,2	Number concept of three	.480	-.522	.369	
M,V,3	Pictorial sim. and diff.	.610	.434	-.225	
M,V,4	Patience:rectangles	.470	.119	-.333	
M,V,5	Comprehension II	.630	.166	.183	
M,V,6	Mutilated pictures	.613	-.187	-.216	
M,VI,1	Number concepts	.663	-.419	.207	.667
M,VI,2	Copying bead chain	.763	-.223	-.015	.495
M,VI,3	Differences	.659	.146	.142	.498
M,VI,4	Response to pictures I	.525	.215	.227	.511
M,VI,5	Counting 13 pennies	.671	-.183	-.127	.528
M,VI,6	Opposite analogies I	.619	.187	.178	.708
	$\Sigma k^2/n$	.365	.079	.059	

TABLE 36

FACTOR LOADINGS FOR ANALYSIS AT AGE 6

Location	Name of test	k_1	k_2	k_3	$k_1(7)$
L,VI,1	Vocabulary	.590	.267	.109	
L,VI,2	Copying bead chain mem. I	.652	-.201	.244	
L,VI,3	Mutilated pictures	.406	.230	.341	
L,VI,4	Number concepts	.770	-.393	-.141	
L,VI,5	Pictorial like. and diff.	.726	.063	-.076	
L,VI,6	Maze tracing	.573	.306	-.135	
L,VII,1	Picture absurdities I	.404	.121	.393	.476
L,VII,2	Similarities: 2 things	.530	.280	-.195	.739
L,VII,3	Copying diamond	.575	-.150	.154	.625
L,VII,4	Comprehension III	.607	.185	-.218	.682
L,VII,5	Opposite analogies I	.549	.135	-.097	.654
L,VII,6	Repeating 5 digits	.497	-.168	-.320	.565
M,VI,1	Number concepts	.667	-.392	-.248	
M,VI,2	Copying bead chain	.495	-.145	.258	
M,VI,3	Differences	.498	.241	.262	
M,VI,4	Response to pictures I	.511	.213	.275	
M,VI,5	Counting 13 pennies	.528	-.263	-.096	
M,VI,6	Opposite analogies I	.708	-.043	-.095	
M,VII,1	Giving no. of fingers	.583	-.362	-.083	.586
M,VII,2	Memory for sentences II	.724	.036	-.369	.636
M,VII,3	Picture absurdities I	.614	.197	-.064	.684
M,VII,4	Repeat 3 digits reversed	.715	-.229	.089	.701
M,VII,5	Sentence building I	.604	.447	-.255	.674
M,VII,6	Counting taps	.522	-.316	.180	.493
	$\Sigma k^2/n$	.352	.062	.048	

TABLE 37

FACTOR LOADINGS FOR ANALYSIS AT AGE 7

Location	Name of test	k_1	k_2	k_3	$k_1(9)$
L,VII,1	Picture absurdities I	.476	-.233	.125	
L,VII,2	Similarities: 2 things	.739	-.463	.283	
L,VII,3	Copying diamond	.625	-.079	.133	
L,VII,4	Comprehension III	.682	-.172	-.076	
L,VII,5	Opposite analogies I	.654	.173	.067	
L,VII,6	Repeating 5 digits	.565	.164	-.253	
L,VIII,1	Vocabulary	.649	.272	.177	-
L,VIII,2	Memory for stories	.620	.168	.191	.558
L,VIII,3	Verbal absurdities I	.740	-.140	-.043	.722
L,VIII,4	Similarities and diff.	.752	-.313	.227	.738
L,VIII,5	Comprehension IV	.754	.256	.170	.662
L,VIII,6	Memory for sentences III	.596	-.103	-.373	.487
M,VII,1	Giving no. of fingers	.586	-.328	-.264	
M,VII,2	Memory for sentences II	.636	.025	-.190	
M,VII,3	Picture absurdities I	.684	.169	.138	
M,VII,4	Repeat 3 digits reversed	.701	.364	-.241	
M,VII,5	Sentence building I	.674	.162	-.412	
M,VII,6	Counting taps	.493	-.007	-.329	
M,VIII,1	Comprehension III	.655	.133	.077	.682
M,VIII,2	Similarities: 2 things	.756	-.233	.344	.742
M,VIII,3	Verbal absurdities I	.770	.255	.100	.808
M,VIII,4	Naming days of week	.656	-.266	-.337	.701
M,VIII,5	Problem situations	.620	-.109	.302	.556
M,VIII,6	Opposite analogies II	.691	.357	.182	.444
	$\Sigma k^2/n$	.438	.054	.055	

TABLE 38

FACTOR LOADINGS FOR ANALYSIS AT AGE 9

Location	Name of test	k_1	k_2	k_3	$k_1(11)$
L,VIII,2	Memory for stories	.558	-.189	.312	
L,VIII,3	Verbal absurdities I	.722	.115	.012	
L,VIII,4	Similarities and diff.	.738	.115	-.178	
L,VIII,5	Comprehension IV	.662	.356	.172	
L,VIII,6	Memory for sentences III	.487	-.098	-.192	
L,IX,1	Paper cutting I	.648	.179	.181	
L,IX,2	Verbal absurdities II	.718	.158	.212	
L,IX,3	Memory for designs	.306	-.139	.347	
L,IX,4	Rhymes:new form	.587	-.198	-.124	
L,IX,5	Making change	.631	-.307	-.120	
L,IX,6	Repeat 4 digits reversed	.524	-.324	-.164	
L,X,1	Vocabulary	.740	.184	.170	-
L,X,2	Picture absurdities II	.393	.232	.122	.428
L,X,3	Reading and report	.594	-.220	.107	.643
L,X,4	Finding reasons I	.516	.219	-.025	.409
L,X,5	Word naming	.387	.062	.001	.480
L,X,6	Repeating 6 digits	.534	-.524	-.306	.499
M,VIII,1	Comprehension III	.682	.146	-.219	
M,VIII,2	Similarities: 2 things	.742	-.181	-.107	
M,VIII,3	Verbal absurdities I	.808	.150	-.176	
M,VIII,4	Naming days of week	.701	-.322	.178	
M,VIII,5	Problem situations	.556	.213	.243	
M,VIII,6	Opposite analogies II	.444	.505	-.104	
M,IX,1	Memory for designs I	.481	-.180	.313	
M,IX,2	Dissected sentences I	.718	-.195	-.076	
M,IX,3	Verbal absurdities II	.731	.262	.100	
M,IX,4	Similarities and diff.	.656	.119	-.223	
M,IX,5	Rhymes: old form	.608	.102	-.242	
M,IX,6	Repeat 4 digits reversed	.425	-.268	-.120	
M,X,1	Block counting	.367	.213	.145	.269
M,X,2	Memory for stories I	.519	-.079	.191	.735
M,X,3	Verbal absurdities III	.680	.163	.110	.723
M,X,4	Abstract words I	.600	.052	.037	-
M,X,5	Word naming: animals	.293	.254	-.332	.497
M,X,6	Repeating 6 digits	.365	-.463	-.249	.456
	$\Sigma k^2/n$	.349	.059	.036	

TABLE 39

FACTOR LOADINGS FOR ANALYSIS AT AGE 11

Location	Name of test	k_1	k_2	k_3	$k_1(13)$
L,X,2	Picture absurdities II	.428	.100	-.512	
L,X,3	Reading and report	.643	.157	.194	
L,X,4	Finding reasons I	.409	-.263	.003	
L,X,5	Word naming	.480	-.168	.296	
L,X,6	Repeating 6 digits	.499	-.467	.190	
L,XI,1	Memory for designs	.325	-.045	-.322	
L,XI,2	Verbal absurdities III	.698	.342	-.181	
L,XI,3	Abstract words I	.868	.240	.195	
L,XI,4	Memory for sentences IV	.579	-.217	.189	
L,XI,5	Problem situation	.605	.353	.183	
L,XI,6	Similarities: 3 things	.596	.141	-.229	
L,XII,1	Vocabulary	.843	.360	.292	-
L,XII,2	Verbal absurdities II	.675	.157	-.101	.705
L,XII,3	Response to pict. II	.683	.142	.086	.576
L,XII,4	Repeat 5 digits reversed	.452	-.425	.257	.449
L,XII,5	Abstract words II	.758	.278	.225	-
L,XII,6	Minkus completion	.610	-.113	.062	.585
M,X,1	Block counting	.269	.062	-.252	
M,X,2	Memory for stories I	.735	.160	.117	
M,X,3	Verbal absurdities III	.723	.291	-.114	
M,X,5	Word naming: animals	.497	-.118	-.003	
M,X,6	Repeating 6 digits	.456	-.397	.156	
M,XI,1	Finding reasons	.554	.045	-.157	
M,XI,2	Copying bead chain mem.	.348	-.239	-.305	
M,XI,3	Verbal absurdities II	.820	-.116	-.055	
M,XII,1	Memory for designs II	.697	-.210	-.406	.611
M,XII,3	Minkus completion	.630	.119	.129	.422
M,XII,4	Abstract words I	.827	.246	.119	.800
M,XII,5	Picture absurdities II	.619	-.027	-.249	.702
M,XII,6	Repeat 5 digits reversed	.595	-.404	.186	.371
	$\Sigma k^2/n$	.381	.060	.049	

TABLE 40

FACTOR LOADINGS FOR ANALYSIS AT AGE 13

Location	Name of test	k_1	k_2	k_3	$k_1(15)$
L,XII,2	Verbal absurdities II	.705	.198	.013	
L,XII,3	Response to pict. II	.576	.273	.125	
L,XII,4	Repeat 5 digits reversed	.449	-.270	-.351	
L,XII,6	Minkus completion	.585	.310	-.190	
L,XIII,1	Plan of search	.555	-.250	.337	
L,XIII,2	Memory for words	.677	-.185	-.066	
L,XIII,3	Paper cutting I	.512	-.209	-.129	
L,XIII,4	Problems of fact	.390	.161	-.164	
L,XIII,5	Dissected sentences	.666	.344	-.138	
L,XIII,6	Copying bead chain mem. II	.465	-.229	-.259	
L,XIV,1	Vocabulary	.850	.250	.308	-
L,XIV,2	Induction	.641	-.252	-.025	.647
L,XIV,3	Picture absurdities III	.674	-.246	.345	.612
L,XIV,4	Ingenuity	.658	-.286	-.045	.713
L,XIV,5	Orientation: direction I	.662	-.239	-.258	.576
L,XIV,6	Abstract words II	.827	.344	.178	.817
M,XII,1	Memory for designs II	.611	-.249	.130	
M,XII,3	Minkus completion	.422	.257	-.254	
M,XII,4	Abstract words I	.800	.400	.213	
M,XII,5	Picture absurdities II	.702	-.236	.222	
M,XII,6	Repeat 5 digits reversed	.371	.144	-.265	
M,XIII,1	Plan of search	.533	-.273	.357	
M,XIII,2	Memory for stories II	.405	.140	.089	
M,XIII,4	Abstract words II	.838	.082	.158	
M,XIII,6	Memory for sentences IV	.522	-.004	-.171	
M,XIV,1	Reasoning	.501	.105	-.121	.673
M,XIV,3	Orientation: direction I	.617	-.169	-.213	-
M,XIV,4	Abstract words III	.857	.223	.171	-
M,XIV,5	Ingenuity	.553	-.296	.092	-
M,XIV,6	Reconciliation opposites	.452	.100	.119	.525
	$\Sigma k^2/n$	.383	.057	.048	

TABLE 41

FACTOR LOADINGS FOR ANALYSIS AT AGE 15

Location	Name of test	k_1	k_2	k_3	$k_1(18)$
L,XIV,2	Induction	.647	-.352	.162	
L,XIV,3	Picture absurdities III	.612	-.160	-.202	
L,XIV,4	Ingenuity	.713	-.231	.044	
L,XIV,5	Orientation: direction I	.576	-.419	-.105	
L,XIV,6	Abstract words II	.817	-.070	-.141	
L,AA,1	Vocabulary	.848	.315	.304	
L,AA,2	Codes	.621	-.079	-.214	
L,AA,3	Diff. abstract words	.901	.128	-.072	
L,AA,4	Arithmetical reasoning	.746	-.130	-.119	
L,AA,5	Proverbs I	.815	.352	-.347	
L,AA,7	Memory for sentences V	.702	.099	.067	
L,AA,8	Reconciliation opposites	.600	-.111	-.166	
L,SAI,2	Enclosed box problem	.407	-.255	.288	.517
L,SAI,3	Minkus completion	.800	.308	-.067	.703
L,SAI,4	Repeat 6 digits reversed	.574	-.144	.430	.592
L,SAI,5	Sentence building	.696	.276	.418	.657
L,SAI,6	Essential similarities	.631	.166	-.244	.595
M,XIV,1	Reasoning	.673	-.176	.365	
M,XIV,6	Reconciliation opposites	.525	-.276	-.167	
M,AA,1	Abstract words III	.864	.106	.213	
M,AA,2	Ingenuity	.747	-.391	-.057	
M,AA,3	Opposite analogies III	.741	.317	.112	
M,AA,4	Codes I	.619	.190	-.078	
M,AA,5	Proverbs I	.798	.457	-.196	
M,AA,6	Orientation: direction I	.698	-.477	-.220	
M,AA,7	Essential differences	.864	.153	-.148	
M,AA,8	Binet paper cutting	.617	-.176	-.243	
M,SAI,1	Minkus completion	.655	.127	-.187	.575
M,SAI,2	Opposite analogies IV	.588	.333	.077	.589
M,SAI,5	Sentence building II	.798	.152	.119	.806
	$\Sigma k^2/n$	.498	.067	.046	

TABLE 42

FACTOR LOADINGS FOR ANALYSIS AT AGE 18

Location	Name of test	k_1	k_2	k_3
L,SAI,1	Vocabulary	.906	.152	.251
L,SAI,2	Enclosed box problem	.517	-.178	-.135
L,SAI,3	Minkus completion	.703	.097	.032
L,SAI,4	Repeat 6 digits reversed	.592	-.292	.090
L,SAI,5	Sentence building	.657	.259	.336
L,SAI,6	Essential similarities	.595	.246	.247
L,SAII,2	Finding reasons II	.673	.031	-.270
L,SAII,3	Repeating 8 digits	.526	-.655	-.043
L,SAII,4	Proverbs II	.692	.231	-.239
L,SAII,5	Reconciliation opposites	.551	.470	-.345
L,SAII,6	Repeat tho't passage	.642	.223	.302
L,SAIII,2	Orientation: direction II	.489	-.021	.271
L,SAIII,3	Opposite analogies II	.630	.127	-.263
L,SAIII,4	Paper cutting II	.475	-.207	.393
L,SAIII,5	Reasoning	.540	.024	-.446
L,SAIII,6	Repeating 9 digits	.656	-.566	-.166
M,SAI,1	Minkus completion	.575	-.014	-.192
M,SAI,2	Opposite analogies IV	.589	.007	.105
M,SAI,5	Sentence building II	.806	.386	.184
M,SAI,6	Reconciliation opposites	.567	.279	-.358
M,SAII,1	Proverbs II	.674	.231	-.268
M,SAII,2	Ingenuity	.448	-.329	-.187
M,SAII,3	Essential differences	.648	.353	.105
M,SAII,4	Repeating 8 digits	.563	-.626	-.232
M,SAII,5	Codes II	.495	-.108	.048
M,SAIII,1	Proverbs III	.668	.400	.139
M,SAIII,2	Memory for sentences V	.367	-.181	.196
M,SAIII,3	Orientation: direction II	.510	.065	.427
M,SAIII,4	Repeating 9 digits	.488	-.699	-.159
M,SAIII,6	Repeat tho't passage II	.739	.437	.068
	$\Sigma k^2/n$	.371	.107	.059

TABLE 43

FIRST FACTOR LOADINGS (AVERAGES) FOR OVERLAPPING AND RECURRING TESTS AND FOR RECURRING TEST SITUATIONS

Experimental age	2	2½	3	3½	4	4½	5	6	7	9	11	13	15	18
Obey. simple commands	.61	.59	.58	.78										
Picture vocabulary	.76	.76	.77	.69	.68	.66	.67							
Identify obj. by use	.73	.66	.53	.69										
Repeating digits	.69	.70	.70		.57	.64		.50	.56	.45	.48			.56
String. beads	.66	.44	.48	.24	.40									
Patience: pictures			.65	.54	.56	.55								
Response to pictures			.75	.66			.52	.51			.68	.58		
Comprehension			.63	.72	.78	.71	.63	.61	.70	.67				
Discrim. animal pict.			.60	.67	.50	.44								
Picture completion: man				.54	.50	.63	.43							
Opposite analogies				.70	.77	.80	.62	.63	.67	.44			.66	.61
Memory for sentences				.66	.67	.57	.57	.72	.62	.49	.58	.52	.70	.37
Pictorial like. and diff.					.47	.58	.61	.73						
Repeat. digits reversed								.72	.70	.48	.52	.41	.57	.59
Vocabulary								.59	.65	.74	.84	.85	.85	.91
Picture absurdities								.51	.58	.39	.52	.69	.61	
Verbal absurdities									.76	.73	.73	.70		
Memory for stories									.62	.54	.73	.40		
Abstract words										.60	.82	.83	.84	
Minkus completion											.62	.50	.73	.64

Chapter X

SPECIAL SCALES

It is the purpose of this chapter to describe, and set forth data on, three special scales: vocabulary as a very abbreviated measure of intelligence, a series of items as a non-verbal scale of intelligence, and a scale for 'immediate memory.' The reason for presenting data on vocabulary should be obvious: there are times when an individually administered, quickly and easily determinable, rough measure of intelligence is needed. The history of intelligence testing provides ample reasons for believing that a suggested non-verbal scale with adequate norms is highly desirable. The frequent use by clinicians of 'memory' items as evidence for or against possible memory deterioration suggests the desirability of two things: first, a critical examination of the meaning and dependability of such 'memory' measures; and second, the presentation of norms so that those who by choice or necessity insist on so gauging memory will at least have an adequate basis for interpreting an obtained score.

Vocabulary

It will be recalled that the vocabulary test consists of 45 words carefully chosen and arranged for difficulty. As a single test it has all the advantages claimed for an individually administered test, and therefore may be preferred to the typical synonym subtest of group scales. Scored by the standards for passing at a given C.A. level, vocabulary tends to yield the highest biserial r's with total score, and scored in terms of number of words passed it yields product-moment correlations with composite M.A.'s of .71, .83, .86, and .83 for ages 8, 11,

14, and 18 respectively (N's of 200 plus at 8, 11, and 14; 101 at 18). These correlations are in part spurious because the vocabulary test is included in M.A. determination, but the degree of spuriousness is not serious since the vocabulary test represents less than 5 per cent of the total number of items entering into M.A. scores. The magnitude of these correlations indicates that the vocabulary test alone constitutes a good rough measure of intelligence. We have no reliability coefficient for the vocabulary test, but the size of its 'validity' coefficients, given above, is such that one need not worry much about reliability.

In Table 44 will be found normative data for the vocabulary test scored in terms of the number of words passed. The N for each age 15 to 18 is 100 plus, and for ages 7 to 14 it is 200 plus with the exception of age 7, where it is only 194 owing to the fact that the test was omitted for 8 cases because of improper location in the provisional form. The performance of 4 of these 8 subjects was such that one can be confident that their scores would have been zero. Accordingly, the computed mean, 7.15, for age 7 has been corrected downward to the 7.0 reported in the table. The needed adjustment to the standard deviation was not made.

It will be noted that the 'growth' curve for vocabulary as here measured shows slight negative acceleration, and that the variability increases with age. A high score of 38 was attained by two subjects, one at age 17, the other at 18. We do not deem it necessary or desirable to present a table of mental-age equivalents by a regression prediction of M.A. from vocabulary score. Since the distributions are fairly symmetrical, the score means and sigmas in Table 44 should permit a reasonably adequate basis for interpreting a given individual's score. We do not recommend the use of this brief vocabulary test as a measure of intellect, although there may be circumstances when it would constitute a far better indicator than no test at all.

TABLE 44

NORMS: MEANS AND STANDARD DEVIATIONS FOR VOCABULARY TEST, SCORED AS NUMBER OF WORDS PASSED

Age	Mean	S.D.
7	7.0	2.0
8	8.3	2.2
9	10.0	2.7
10	11.4	3.1
11	13.7	3.9
12	15.4	4.7
13	17.4	5.1
14	18.2	4.9
15	20.0	5.2
16	21.2	5.4
17	22.0	4.8
18	22.5	5.6

A Tentative Non-Verbal Scale

When assembling the items for the New Revision, it was hoped that enough non-verbal material could be included to permit the construction of a non-verbal form which would parallel two verbal forms as regards difficulty, reliability, and validity. Despite the large number and diversity of the items utilized in the preliminary work and later in the two provisional forms, it was not possible to realize this goal. It was thought desirable, however, to include as much of the non-verbal material, especially at the lower end, in the final forms as seemed to satisfy the requirements laid down for the retention of items. The presence of this more or less non-verbal material in the final form has led to the question as to what might be expected of these retained items as a separate scale. Accordingly, we here present a brief sta-

tistical analysis of two non-verbal forms, somewhat balanced for content, difficulty, and validity. The 40 items utilized were selected by Dr. Merrill as being non-verbal or at least as being less verbal than the other tests contained in Forms L and M. Since the directions for these items are mainly verbal rather than pantomime, it follows that some understanding of language is involved and consequently that the items are not to be regarded as purely non-verbal.

The 20 items for Form I and for Form II are listed in Tables 45 and 46 by location in Forms L and M and by name. It will be noted that those in Form I are predominantly from Form L, and that 14 of the 20 tests of each form are located at level VI or lower, with only 8 items scattered from level VII to the adult levels. Although we have analyzed the results for each age from 2 to 18, the fewness of items beyond age level VI does not warrant any detailed presentation of data for ages beyond 8. The essential data regarding age means, sigmas, form versus form reliabilities, and correlations of each form with M.A. based on a composite of Forms L and M are presented in Table 47. Because of the restricted range in non-verbal scores, all correlations in this table are tetrachorics. Their standard errors will be about .10 to .12 for ages 2 to 5-1/2, and about .07 to .08 for ages 6, 7, and 8.

Reference to Table 47 shows that the scales correlate moderately with mental age, but these correlations are spuriously high because the non-verbal tests are involved in mental ages. If this spurious element were eliminated, one might expect the correlation to average about .65 instead of about .70. The reliabilities average near .65. One must refrain from correcting for attenuation the correlations between M.A. and the non-verbal scales because the measurement errors will surely be correlated. It would appear that a scale of so few items yields reliabilities of insufficient size to warrant recommending the use of a single form; both forms combined

TABLE 45

ITEMS INCLUDED IN TENTATIVE FORM I OF NON-VERBAL SCALE

L,II,1	Three-hole form board
L,II,4	Block building: tower
L,II-6,6	Three-hole form board: rotated
L,III,1	Stringing beads
L,III,5	Copying a circle
L,III-6, alt.	Drawing a cross
M,III-6, alt.	Matching objects
L,IV,3	Picture completion: man
L,IV,5	Discrimination of forms
M,IV, alt.	Discrimination of animal pictures
L,V,1	Picture completion: man
L,V,2	Paper folding: triangle
L,V,4	Copying a square
L,VI,2	Copying a bead chain from memory I
L,VII,3	Copying a diamond
L,IX,1	Paper cutting I
L,IX,3	Memory for designs
L,XI,1	Memory for designs
L,XIII,6	Copying a bead chain from memory II
L,S.A.III,4	Paper cutting II

TABLE 46

ITEMS INCLUDED IN TENTATIVE FORM II OF NON-VERBAL SCALE

M,II,1	Delayed response
M,II-6,2	Motor coordination
M,II-6, alt.	Stringing beads
M,III,4	Drawing a vertical line
L,III,3	Block building: bridge
L,III, alt.	Three-hole form board: rotated
M,III-6,3	Discrimination of animal pictures
M,III-6,5	Sorting buttons
M,IV,2	Stringing beads
M,IV-6,1	Discrimination of animal pictures
M,IV-6,4	Picture completion: bird
M,V,4	Patience: rectangles
L,V, alt.	Knot
M,VI,2	Copying a bead chain
M,IX,1	Memory for designs I
M,X,1	Block counting
M,XI,2	Copying a bead chain from memory
M,XII,1	Memory for designs II
L,XIII,3	Paper cutting I
M,A.A.,8	Binet paper cutting

TABLE 47

DATA ON NON-VERBAL SCALES

	Form I		Form II		Correlations			Composite or sum of Forms I & II	
Age	M	S.D.	M	S.D.	I vs. M.A.	II vs. M.A.	I vs. II	M	S.D.
2	2.1	1.2	1.9	1.2	.53	.60	.53	4.0	2.1
2½	3.7	1.7	4.4	1.9	.74	.64	.75	8.1	3.4
3	5.6	2.0	6.3	2.1	.71	.84	.56	11.9	3.6
3½	8.0	2.2	8.6	1.9	.71	.64	.76	16.6	3.8
4	9.3	2.3	9.9	1.9	.77	.57	.61	19.2	3.8
4½	11.3	2.1	11.2	1.8	.63	.76	.53	22.5	3.4
5	12.2	1.9	12.4	1.5	.53	.66	.79	24.6	3.2
5½	13.0	1.4	13.1	1.1	.67	.72	.71	26.1	2.3
6	13.7	1.4	13.7	1.0	.77	.61	.76	27.4	2.3
7	14.9	1.5	14.3	1.0	.70	.58	.52	29.2	2.2
8	15.8	1.2	14.9	1.2	.60	.70	.55	30.7	2.1

would have a reliability in the vicinity of .79, which is also unsatisfactory.

The unreported data for age groups 9 to 18 show still lower reliabilities, and also lower correlations for the non-verbal versus mental ages. This, of course, is not unexpected since there are still fewer non-verbal items at these levels. The distributions become more and more skewed for lack of top. It might be of interest to note here that the 40 items chosen as non-verbal tend to have first factor loadings which average .50 as compared to about .60 to .65 for all the items of the New Revision. The fact of variation among the 40 items as regards their general factor saturation suggested that a system of weights based upon magnitude of first factor loadings might improve the correlations between the non-verbal scales and mental age. Actual tryout at ages 4, 5, 10, and 14 gave evidence that such a basis for weighting would not increase the correlations.

Memory

From each of Forms L and M, 22 items, well scattered throughout the age levels, were chosen on an a priori basis as items which could be said to measure 'memory,' or more precisely 'immediate memory.' The tests or items so selected are listed in Tables 48 and 49. Two memory scores were determined for all the individuals in the standardization group from ages 2 to 18. The two sets of memory scores were correlated with each other and each was correlated with composite mental age. These correlations were computed as tetrachorics because of the limited range of the memory scores. Means and standard deviations for the two scales, and for scores obtained by combining the two, were calculated. These statistics, which are presented in Table 50, need little discussion.

The reliabilities tend to average .70, which when

SPECIAL SCALES

TABLE 48

MEMORY SCALE. FORM I (L)

II-6,5	Repeating 2 digits
III,4	Picture memories
III,6	Repeating 3 digits
IV,2	Naming objects from memory
IV-6,2	Repeating 4 digits
IV-6,5	Three commissions
V,5	Memory for sentences II
VI,2	Copying a bead chain from memory I
VII,6	Repeating 5 digits
VIII,2	Memory for stories
VIII,6	*Memory for sentences III
IX,3	Memory for designs
IX,6	Repeating 4 digits reversed
X,3	Reading and report
X,6	Repeating 6 digits
XII,4	Repeating 5 digits reversed
XIII,2	Memory for words
XIII,6	Copying a bead chain from memory II
A.A.,7	Memory for sentences V
S.A. II, 3	Repeating 8 digits
S.A. II, 6	Repeating thought of passage
S.A.III,6	Repeating 9 digits

TABLE 49

MEMORY SCALE. FORM II (M)

II,1	Delayed response
II-6,5	Repeating 2 digits
III,6	Repeating 3 digits
IV,6	Memory for sentences I
IV-6,3	Repeating 4 digits
VII,2	Memory for sentences II
VII,4	Repeating 3 digits reversed
IX,1	Memory for designs I
IX,6	Repeating 4 digits reversed
X,2	Memory for stories I
X,6	Repeating 6 digits
XI,2	Copying a bead chain from memory
XI,6	Memory for sentences III
XII,1	Memory for designs II
XII,6	Repeating 5 digits reversed
XIII,2	Memory for stories II
XIII,6	Memory for sentences IV
S.A. I,4	Repeating 6 digits reversed
S.A. II,4	Repeating 8 digits
S.A. III,2	Memory for sentences V
S.A. III,4	Repeating 9 digits
S.A. III,6	Repeating thought of passage II

TABLE 50

DATA ON MEMORY SCALES

	Form I (L)		Form II (M)		Correlations			Composite or sum of Forms I & II	
Age	M	S.D.	M	S.D.	I vs. M.A.	II vs. M.A.	I vs. II	M	S.D.
2	.8	.9	1.3	.8	.79	.75	.66	2.1	1.5
2½	2.2	1.4	2.2	1.2	.87	.72	.78	4.4	2.6
3	2.8	1.4	2.9	1.3	.60	.62	.59	5.7	2.4
3½	4.5	1.6	3.9	1.0	.83	.56	.69	8.4	2.4
4	5.0	1.6	4.1	.9	.83	.57	.68	9.1	2.3
4½	6.2	1.7	4.6	.8	.83	.83	.78	10.8	2.4
5	6.9	1.7	4.9	.9	.56	.58	.77	11.8	2.5
5½	7.6	1.7	5.2	1.0	.77	.76	.55	12.8	2.4
6	8.2	1.7	5.7	1.2	.67	.92	.65	13.9	2.6
7	10.0	2.1	7.1	1.7	.79	.80	.74	17.1	3.5
8	11.4	2.1	8.7	2.2	.78	.66	.71	20.1	4.0
9	13.0	2.3	10.4	2.5	.77	.75	.80	23.4	4.6
10	14.3	2.2	12.0	2.6	.82	.86	.68	26.3	4.4
11	15.3	2.3	13.1	3.0	.75	.83	.78	28.4	5.0
12	16.0	2.5	14.2	2.9	.68	.84	.68	30.2	5.0
13	16.9	2.2	15.2	2.4	.76	.80	.74	32.1	4.2
14	17.0	2.0	15.5	2.3	.82	.83	.82	32.5	4.1
15	17.5	2.1	16.1	2.8	.90	.88	.69	33.6	4.5
16	17.9	2.2	16.4	2.6	.74	.77	.80	34.3	4.6
17	18.2	1.9	16.8	1.9	.71	.56	.54	35.0	3.3
18	18.3	2.0	16.9	2.2	.86	.75	.70	35.2	3.9

stepped up would indicate a reliability of about .82 for scores based on the items in both scales. Since clinicians and other workers seldom administer more than one form, the memory ability which they infer from the 'memory' items will possess a reliability of only .70, which is too low for individual diagnosis.

To argue that such an unreliable measure of memory is better than none at all overlooks another pertinent fact: the memory scores correlate just about as high with mental age as the reliabilities permit. An exact determination of the correlation to be expected between perfectly measured mental age and memory (as here defined operationally) is complicated by the spurious nature of the obtained correlations between memory and mental age and by the likelihood of correlated errors. Any reasonable allowance for these effects will lead to the conclusion that 'memory' as determined by the items of a 'memory' nature in the New Revision is not very different from the general intelligence being measured by the scale as a whole. The first factor loadings for the memory items average about .05 lower than the average for all the items; thus the memory items are not quite so highly saturated with the central function being measured. It would appear, therefore, that those clinicians who continue to have faith in the utility of certain Binet items as a measure of memory or 'immediate' memory may find but little to support their position.

Of course, this whole issue is a problem in a broader field of study, namely, the organization of abilities. Our own factor analyses (see Chapter IX) are too limited to throw much light on 'memory' as a factor. Items which logically seem to call for some sort of memory do not have similar factorial patterns for the three factors extracted. It is true, however, that the repeating-of-digits tests do possess similar loadings, but whether the fact indicates more than a specific repeating-of-digits factor is questionable. Perhaps a thoroughly intensive and extensive factor analysis of memory,

based upon a large sample or samples, is needed. But such an analysis might not provide an answer to those who claim that intelligence tests of the Binet type are merely measures of memory. The argument assumes that two individuals scoring different mental ages differ primarily as regards memory ability, i.e. power of retention and reproduction. Such a concept would rule out individual differences in the ability to observe, see relationship, or profit by experience. After all, observation and learning must precede retentivity; even in the case of repeating digits we have an example of single-trial learning. The final answer as to whether variance in measured intelligence is more dependent upon retentivity than upon original learning, or as to the extent of each as a contributor, must be sought in the laboratory. We hazard the guess that securing the answer will involve experimentation rather than wholesale correlational analysis.

Summary

The materials of this chapter have been presented in order to emphasize the limitations of special scales made up by selecting appropriate items from Forms L and M. The vocabulary test alone yields a fairly adequate measure of the kind of intelligence measured by the New Revision. The dearth of non-verbal material is such that little reliance can be placed upon a score based solely upon the non-verbal items. Regarding 'memory' as inferred from items which apparently involve memory or immediate memory, we have called attention to the low reliability of such scores and have questioned the logic of assuming that memory, as a function, is really the trait tapped by these items.

In view of the questionable validity of the 'memory' items as a scale, in view of the findings of the factor analyses, and in view of the low reliability for single

items, we find ourselves in perfect agreement with Goodenough,[1] who very aptly states her opinion that 'a test of general intelligence cannot be made to serve the purpose of a universal diagnostic instrument', and that 'the practice, so unfortunately common among clinicians, of making pronouncements about special abilities or defects in such broad psychological categories as memory, visual imagery, perception, and the like on the basis of one or two items in a Binet test, is hazardous in the extreme.'

[1] F. L. Goodenough, "Review of *Measuring Intelligence*," *Psychol. Bull.*, 1937, 34, 605-609.

Chapter XI

UNITS OF MEASUREMENT

Much has been written in the field of psychology and education about measurement, and one of the topics of chief interest and controversy has been the units proposed. Many of the units used are purely arbitrary, even to the extent of being accidental. So far as we know, no one has brought forth a scale based upon units which would satisfy the criteria of equality used in the physical sciences. It does not follow from this that psychological measurement is impossible unless one restrict the word 'measurement' to situations wherein a scale possessing physically equal units is employed. Not all the scales used in the physical sciences have equal units; this fact does not nullify, but does restrict, their use.

Psychologists must at present be content to utilize scales which have limitations. At least one claim can be made for nearly, if not all, psychological scales: they permit the rank ordering of individuals, subject, of course, to an ever present and at times disturbingly large error. Now, the choice of unit must depend partly upon preference and partly upon general usefulness. In *Measuring Intelligence* reasons were stated for retaining the M.A. and I.Q. scheme of scoring. We are not blind as to the shortcomings of such a system of units. Not only have no claims been made for the equality of mental age units but actually their inequality has been admitted (see pages 24-29 of *Measuring Intelligence*). In fact, the use of I.Q. units is predicated on inequality of mental age units.

It is not our purpose here to review the reasons for perpetuating the mental age and I.Q. concepts. We do propose to examine some of the alternatives with the special intent of scrutinizing the merits claimed for them.

Before doing this we should like to digress long enough to discuss some of the criticisms set forth in a recent paper by Richardson.[1]

Richardson's Logic About Age Scales

We have pointed out in Chapter VIII on per cents passing items that Richardson seems to have been misled as to the method of placing items in age levels. He evidently believes that the sole criterion of item validity used was the steepness of the curves for per cent passing, i.e. the item's correlation with age, although it is clearly stated by Terman and Merrill that 'the correlation of each test with composite total score (equivalent to correlation with mental age) was computed separately for each test, thus providing a basis for the elimination of the least valid tests,' (*Measuring Intelligence*, page 22.) He also thinks that nothing in the procedure used operates so as to select items that measure a unique trait. These notions are so erroneous as to need no comment.

As an example of logic brought to bear upon a 'logical difficulty,' we quote: 'Let us assume further that two items are so discriminating and so far apart in proper age-location that their discrimination functions do not overlap. The result is that two hypothetical "good" items or sub-tests have a zero correlation. A scale made up of such items must necessarily be unreliable as a composite.' If this reasoning is correct and if the conclusion therefrom has any meaning at all, it would follow that no scale could be reliable which contained items differing markedly as to difficulty, and that the possibility of constructing a scale applicable over several ages or grades would be ruled out. A rote-memory measure like repeating two, three, four ... ten digits would not

[1]M. W. Richardson, "The Logic of Age Scales," *Educ. Psychol. Measmt.*, 1941, 1, 25-34.

be admissable. One would like to know how Richardson would go about measuring the intellect of a 6-year-old and of a 16-year-old in a manner that would avoid his 'logical difficulty.'

On page 25 of the Terman and Merrill volume it was stated that the expression of a test result in terms of age norms rests on no statistical assumptions. This is characterized by Richardson as 'erroneous and misleading,' and he goes on to say, 'The truth of the matter is that *mental age* is a measure derived from raw scores in accordance with certain assumptions. It is regrettable that he did not specify these assumptions.

The greatest confusion in Richardson's paper is to be found in his discussion of I.Q. constancy. It is well known that there are several necessary conditions for I.Q. constancy. Some of these are mentioned by Richardson, but treated as sufficient conditions. A few quotations are in order: 'The constancy of the I.Q., if it exists, is imposed by the process of standardization.' 'If the various sub-tests are properly scaled ... the I.Q. of 100 will remain constant.' 'The gist of the matter is that the I.Q. can be made to be constant.'

All the statements just quoted are false. Let us list the conditions necessary for I.Q. constancy, i.e. for an individual obtaining the same I.Q. within error limits, on successive testings over a period of time, e.g. 2 to 13 years. (1) The same general intellectual ability must be called for at the various age levels. (2) The standard deviation of successive age I.Q. distributions must be equal (this means a systematic increase in the sigma for M.A. distributions.) (3) I.Q.'s and age must be uncorrelated. These conditions, which can be reasonably well attained, are functions of the scale, and even if perfectly achieved neither they nor any other scale functions, will guarantee I.Q. constancy. In other words, they are not sufficient conditions. In order to find the latter, we must look to the individual. If the above necessary conditions obtain, then a sufficient, also necessary, condition for

I.Q. constancy is that the growth rates of all individuals remain constant. Few will agree that the question of growth rate is a 'false issue' as claimed by Richardson. According to him the constancy of the I.Q. is an example of a problem where 'the distinction between purely psychometric issues and psychological issues [is] not always made.' We find ourselves wondering why he himself did not make the distinction instead of predicating that the whole thing is a psychometric problem.

Richardson's notion regarding the cause of increase in M.A. variability with age is interesting. He states that 'we may increase the standard deviations of successive year levels by (a) selecting sub-tests which have higher intercorrelations at older age levels, (b) assigning a larger number of mental months to each subtest.' This latter scheme is said to be inadmissible; hence 'the conclusion is inescapable that the degree of correlation between sub-tests must increase steadily with higher age levels if the I.Q. is to be constant.' It happens that we can present some correlations which are more inescapable than the outcome of Richardson's logical argument. The average intercorrelations between items beginning with age 2 are as follows: .44 (2), .35 (2-1/2), .36 (3), .41 (3-1/2), .40 (4), .35 (4-1/2), .36 (5), .34 (6), .43 (7), .33 (9), .36 (11), .36 (13), .48 (15), and .36 (18).

One more example of this author's logic: 'If half-year groups have the same I.Q. dispersion, they must have approximately the same mental age dispersion. But the mental age dispersions must increase from year to year in order for I.Q.'s of individuals to be constant. It thus appears that two possible properties of the I.Q. are inconsistent, and not attainable at the same time, in any strict sense.' This is an example of a false premise leading to a false conclusion — the 'if' statement is simply untrue.

Note on Standard and T-Scores

Many psychometricians have urged the universal adoption of some form of the standard-score unit. We are well aware of certain advantages which would accrue therefrom, but we are not convinced that these advantages outweigh the interpretative value of the mental age-I.Q. combination. Furthermore, we are unable to appreciate some of the properties claimed for units of the standard-score variety. This section will be devoted to consideration of some of those claims.

The standard score is perhaps better adapted to the needs of researchers. Presumably the advantage is primarily that of having all tests scored in the same type of unit, hence making for greater comparability than is possible with the arbitrary and often accidental point scores. In particular it has been argued by some that the use of standard scores would per se avoid the problem of such differences in I.Q.'s as are found when one passes from scale to scale. But in order to be sure that such differences are really eliminated, or that the standard scores are really comparable, one needs to satisfy the condition that the several tests shall have been standardized on samplings which are comparable as regards general level and scatter of ability.

The claim by a surprisingly large number of psychologists that the use of the standard-score method will yield units which are equal or 'truly' equal deserves some attention. At this point we should distinguish between two variant methods for deriving a score of the standard type. First, there is the relatively simple scheme of dividing deviations from the mean by the standard deviation of the distribution. This, for some writers, is the accepted technique for obtaining standard or z-scores. The introduction of a constant multiplier, say 10, and an additive constant, say 50, will so transform z-scores as to set the mean at 50 and the sigma as 10, thereby getting rid of negative scores and permitting the

elimination of decimals. The resulting scores have sometimes been called T-scores, but it would seem wise to restrict the term T-score or T-scale to its original meaning, which is associated with the method of scaling used. This method, which constitutes our second variant, depends upon converting the per cent attaining a given raw score into an equivalent sigma by use of the normal curve functions.[1] The resulting score is so adjusted as to yield a mean of 50 and a sigma of 10, but it does not follow from this that 'T-scores are σ-scores [or z-scores] multiplied by 10 ....' This relationship holds only when the original distribution of raw scores is normal. In the discussion to follow, the reader will do well to keep in mind the distinction which we have made between z-scores and T-scores, i.e. between the units which result from dividing by sigma and those which are derived by recourse to normal curve functions.

Let us now examine the claim that z-scores (also T-scores) are 'truly equal' units and can be treated as 'though they were all in inches or pounds.' First, consider the z-score. By definition we have

$$z = \frac{X-M}{\sigma} = \frac{X}{\sigma} - \frac{M}{\sigma}$$

which is obviously a linear relationship of the form $Y = BX + A$; hence if the original X-units are equal, the transformation will yield z-units which are also equal. But suppose either that the original X-units were unequal or that we were ignorant as to their equality; will a simple linear transformation make equal units out of unequal units, or will our ignorance be metamorphosed into knowledge by such simple arithmetic? The answer is so obvious that the question should hardly be necessary.

Next, let us consider the T-score. One thing is accomplished by T-scaling which is not achieved by the

[1]For detailed explanation see pages 151-157 of H.E. Garrett, *Statistics in Psychology and Education.* New York: Longmans Green Company, 1937.

z-transformation; namely, the distribution of T-scores will be normal, at least for the sample used in the scaling, while that for z-scores will have the same shape as the distribution of original raw scores. We are aware of the advantages of normal score distributions — our concern here is whether or not the normalizing by T-scaling has resulted in a scale of 'truly equal' units. Is there evidence that such is the case, or is it merely assumed? We know of no supporting evidence, hence we question the tenability of the assumption. As we are frankly skeptical about a purely logical approach to the problem, we resort to an example.

A distribution containing measurements on 7749 men has been T-scaled by the author. This N is sufficiently large to permit of fairly exact scaling, particularly for points near the center of the distribution. It is found that the difference between 130 and 140 original units corresponds to 5.5 T-units, while the difference between 190 and 200 is equivalent to 3.4 T-units. Thus, if T-units are equal, it means that a 10-point difference in original units in one region is not equal to a 10-point interval in another region; in fact, one is 60 per cent larger than the other. If this is true, we have proved that the difference between 130 and 140 pounds is 60 per cent larger than the difference between 190 and 200 pounds. If, starting with a scale of truly equal units (pounds), one comes out with T-units which are definitely unequal, what can be expected when one starts with a scale of arbitrary, admittedly unequal units?

Another type of unit has been frequently used, namely that which results from the so-called absolute scaling methods. In this instance the original data are the per cents passing items rather than a distribution of scores. So far it has not been demonstrated that absolute scaling leads to equal units. There is at least one reason why one cannot expect any of the scaling methods (absolute or standard or T) to yield units which are 'truly equal,' viz. the fact that all scaling must be

done on data based upon a sample of individuals. The sampling errors — in the sigma of a distribution or in the per cent exceeding a score or in the per cent passing an item — will always be present, and consequently the derived units will be subject to chance fluctuations. It follows, therefore, that these methods cannot possibly be expected to yield equal units. We do not conclude from this that such units have no desirable properties — we have merely refuted the somewhat generally accepted claim that scaling leads to 'truly equal' units.

Heinis Mental Growth Units

The growth curve and unit of growth proposed by Heinis[1] have received considerable attention among certain workers. We are not primarily concerned about the adequacy of the Heinis growth curve — it may describe mental development as accurately as any of the proposed curves, but it was originally deduced from such small samples (with little information as to the nature of the samplings) that one might rightfully doubt its generality. We, personally, believe that the exact form of the mental growth curve is unknown and will likely remain unknown. So far as mental measurement is concerned, the chief issues center about the relative merits of various methods of expressing scores. In particular, we should like here to examine the claim, recently accepted by Kuhlmann, that the Heinis personal constant (P.C.) is more constant than the I.Q.

Strictly speaking, the issue has to do with one of the conditions necessary for constancy, namely that the variability of score distributions be the same, or nearly so, for successive age groups. The question, therefore, is whether I.Q. scoring or P.C. scoring yields the great-

[1] H.A. Heinis, "A Personal Constant," *J. Educ. Psychol.*, 1926, 17, 163-186.

er consistency as regards variability measures. Kuhlmann[1] has made two comparisons from which he concludes that P.C.'s are (more exactly, can be) more constant than I.Q.'s. His first comparison is based on the P.E. (defined as $Q_3 - Q_1$ instead of the usual $Q_3 - Q_1$ divided by two) of I.Q. and of P.C. for his own test scored by both methods. The I.Q. variabilities fluctuate more, and are particularly high for the upper age levels, as might be expected when a straight-line growth curve is assumed as far as age 16. But this comparison may not be a fair one, since his scale was constructed on the basis of the Heinis curve, and therefore the P.C.'s might be expected to make a better showing.

Kuhlmann's second comparison is somewhat more crucial in that the variabilities for P.C.'s based on scaling according to the Heinis curve are contrasted with the variabilities for I.Q.'s based upon the new Stanford revision. The Kuhlmann and the new Stanford-Binet may be thought of as the best tests yet constructed on the basis of their respective underlying assumptions. It is doubtful whether either test approaches perfection, and it is known that there are real differences in variabilities for the new Stanford-Binet. The question is whether the fluctuation in standard deviations from age to age for I.Q.'s (standardization data of the new Stanford-Binet revision) is greater than the variation of P.E.'s for P.C.'s (Kuhlmann's data). Kuhlmann's chief deductions are based on a consideration of data for ages 6 to 16; therefore we will limit our discussion to these ages.

In Table 51 will be found the S.D.'s (mean values for Forms L and M) for I.Q. distributions and the P.E.'s (interquartile range) for P.C.'s. The latter values come from Table XXVIII of Kuhlmann. The former have been rounded off so that they will not be reported any more exactly than the available P.E.'s for the personal constant. Kuhlmann pictures these values in his Fig. I,

[1]F. Kuhlmann, *Tests of Mental Development.* Minneapolis:Educational Test Bureau, 1939; see pages 86-93.

page 91, and concludes that I.Q.'s are less constant in variation than are P.C.'s. He believes that the 'evidence is conclusive as to the relative merits of I.Q. and P.A. [i.e., P.C.] scores.' The P.C. remains 'quite constant at *all* levels above six ... whereas the I.Q. does change at varying rates ...' These conclusions were evidently based upon the fact that the range of S.D.'s for I.Q.'s is from 13 to 20 as opposed to the range for P.E.'s of P.C.'s, which is from 6 to 9. If we deal with averages, we find that the former variabilities center about a mean of 17 and yield an average deviation of 1.27 and an S.D. of 1.76, whereas the mean for the latter is 8.1 with an average deviation of .83 and a sigma of 1.01. Insofar as constancy depends upon stable variabilities, it would seem that the P.C. would permit greater constancy, and that Kuhlmann was correct.

But if we examine the data in a different manner, a different and more valid conclusion will emerge. This is a situation wherein the variation of measures of variability must be considered in a relative sense in order to make a proper allowance for the fact that we are dealing with measures based on noncomparable units. One would not be justified in concluding that the real variation in intelligence is greater when measured in

TABLE 51

COMPARISON OF FLUCTUATIONS OF VARIABILITY MEASURES FOR I.Q.'S AND P.C.'S: S.D. FOR I.Q.'S AND P.E. (INTERQUARTILE, NOT SEMI-INTERQUARTILE RANGE) FOR P.C.'S

Age	6	7	8	9	10	11	12	13	14	15	16
$S.D._{I.Q.}$	13	16	16	17	16	18	20	18	17	19	17
$P.E._{P.C.}$	9	8	7	9	7	9	8	8	9	6	9

terms of I.Q. than when gauged in P.C.'s simply because the S.D. for a distribution of I.Q.'s is numerically greater than the S.D. for a distribution of P.C.'s. It is obvious that the variability of I.Q.'s is numerically larger than that for P.C.'s; hence when comparing the variation of the measures of the variabilities, we must take into account this difference in absolute value which is a function of the original unit used. In other words we cannot compare directly the average deviations of .83 and 1.27, or the sigmas of 1.01 and 1.76, any more than one can compare numerical, untransformed values based upon inches and centimeters. When we take .83 and 1.27, or 1.01 and 1.76 as measures of variation relative to the respective means, 8.1 and 17, we see at once that the variation of the P.E.'s for the personal constant is in reality greater than the variation of S.D.'s for the intelligence quotient. There can be no objection to using the coefficient of variability in this case, since the means, 8.1 and 17, can be thought of as distances above real zero points, i.e. the points of no variability.

Another way of considering the figures in Table 51 is to examine the range of variabilities. Thus for I.Q. scoring the S.D.'s range from 13 to 20, the larger being 54 per cent greater than the smaller; for P.C. scoring the P.E.'s range from 6 to 9, the latter being 50 per cent greater than the former. The problem may be viewed in still another manner. If the I.Q. rating of an individual on the new Stanford-Binet did remain constant, a score of 113 at age 6 would correspond to the 84th percentile, while a score of 113 at age 12 would fall at the 74th percentile. If the P.C. rating of an individual on the Kuhlmann test did remain constant, a score of 103 at age 15 would be the equivalent of the 75th percentile, and the same score at age 16 would be near the 66th percentile. These shifts in percentile ratings, which represent the maximum to be expected in the case of each scoring scheme, are so similar that it is impossible to adjudge that one method fares better than the other.

It would appear, therefore, that the relative merits of mental ages and I.Q.'s versus Heinis mental units and P.C.'s cannot be decided on the basis of the difference in the success with which their respective advocates have met one of the conditions necessary for constancy of indices of intellectual brightness.

Chapter XII

SUMMARY

The materials which have been presented herein are complementary and supplementary to Terman and Merrill's *Measuring Intelligence*. Some chapters contain more detail on certain topics than was feasible to give in the previous volume, while other chapters present data not heretofore reported. Some parts of the present work are devoted to data involving total scores; other parts are concerned with data on items. Any summary of such a large mass of data is indeed difficult, since brevity means the omission of necessary qualifications. It is hoped, however, that the following resume will serve a useful purpose.

I.Q. Distributions. - The distributions of I.Q.'s for Forms L and M tend to approach the normal curve type. The discrepancies, although in some instances statistically significant, are nevertheless small in magnitude. No conclusions regarding the distribution of intellect were drawn; as measured by the new Stanford-Binet, intelligence is for all practical purposes distributed in the normal fashion.

I.Q.'s and School Progress. - The analysis of I.Q.'s by age-grade location provides interesting information about intellect as related to school progress. The modal age-grade (normal-progress) groups are average in I.Q., while those accelerated or retarded by one grade have I.Q.'s which average about 11 points above or below the general average. Those who are two grades ahead or behind deviate about 22 I.Q. points from the averages for the modal age groups. There seems to be no difference in I.Q. variability for age groups as opposed to grade groups, but as regards mental ages the grade groups are somewhat more homogeneous than age groups. Within a grade

group, however, one finds a rather wide range of mental ages.

I.Q. by Occupation and Residence. - The data on I.Q. differences associated with occupational levels and with urban, suburban, and rural residence tend to confirm previous findings. Such differences as exist emerge at the early ages and continue throughout the age range here utilized.

Sibling Resemblances. - The sibling resemblances reported are of particular interest because of the fact that highly reliable I.Q.'s (average of Forms L and M) are involved and because all the resemblance coefficients are based on groups of typical I.Q. variation. For siblings of all ages, 2 to 18, the coefficient for 384 pairs is .53; for 42 pairs of preschool age, .55; and for 119 pairs, one of each pair being of preschool age, the older being between ages 6 to 18, the resemblance is .52.

Sex Differences. – Sex differences in I.Q. tend to be small – about 3 points in favor of girls at the preschool ages and about 2 points in favor of boys at the later ages. These differences, although not large, are near the borderline of statistical significance. Some of the eliminated and also some of the retained items yielded highly significant differences in per cents passing. It was suggested that the student of sex difference might profitably direct his attention toward items rather than total scores, which may mask real differences between the sexes.

Reliability. - Chapter VI contains a detailed analysis of reliability. It is shown that the standard error of measurement for the I.Q. is definitely related to the magnitude of the I.Q. The size of the measurement errors as well as the equivalent reliability coefficients were deduced by averaging the results obtained by two different methods. The standard errors of measurement, as determined for ages 6 to 13, range from 2.8 for low I.Q.'s to 5.3 for I.Q.'s in the higher brackets, and the equivalent reliability coefficients range from .97 down to .90, the higher reliability being associated with the lower

I.Q.'s. Thus it is no longer permissible to speak of 'the' reliability of a scale of the Binet type, and it is likely that the same thing is true for group tests which utilize the mental age-I.Q. concepts.

Scatter. - The spread of individual performance, or scattering of successes and failures, was discussed in Chapter VII. Scatter is, of course, the result of several different factors: item unreliability, low intercorrelations among items, lack of high correlation of items with age (these are highest for items at the lower end of the scale), the presence of a series of items which call for some special ability, and lastly faulty age placement of items. These are all functions of the scale; the first three are inescapable, while the last two can be, and in the New Revision have been, fairly well eliminated. One feature has definitely contributed to scatter, viz. the presence of recurring tests and recurring test situations. A reason for an apparently greater scatter on the new Stanford-Binet, as compared to the 1916 Revision, is the inclusion of items at additional age levels. Scatter may also be a function of chance motivational factors in the testee, but since we have shown that the form versus form reliability of scatter scores approaches the vanishing point, it becomes difficult to see how any clinical meaning can be attached to the concept of scatter.

Per Cents Passing. - Table 26 of Chapter VIII contains the per cents passing items by age. The corresponding curves are steepest for the items at year level II, and become less steep for higher levels. They are skewed rather than normal ogives. Although the curves for items located at a particular age level do not cross the ordinate for that age at the 50 per cent point of difficulty, there are items in the scale which are of medium difficulty for each age, except age 6. This lack of items of medium difficulty at this age was offered as an explanation for the low observed I.Q. variability for that age group. Reasons were given for believing that our data on per cents passing may not be ideal as basic

data for deriving a mental growth curve.

Factor Analyses. - The fourteen factor analyses reported were arranged so as to include each item in at least one analysis and to provide overlapping items between adjacent analyses. The results indicate that the several items included in a given analysis tend to be saturated, though in varying degrees, with a common factor. This was not surprising since the methods used in selecting items were such as to favor this outcome. There is some evidence for minor group factors, but the sampling errors are so large as to make difficult any very specific deductions concerning these factors. It would appear that the items in the scale are not measuring such a hodgepodge of abilities as some have supposed. Presumptive evidence was presented to the effect that the common factors at successive levels are nearly identical. Insofar as this is true and insofar as the amount of variance due to possible group factors is small, we have evidence that I.Q.'s earned on the New Revision are comparable quantitatively and qualitatively.

Special Scales. - The materials on special scales have been presented with the idea of drawing attention to the inadequacies of scores based on so few items. The vocabulary test does constitute a fair measure of the type of general intelligence measured by the Binet test as a whole, but of course one cannot conclude from this that intelligence depends upon vocabulary ability rather than the converse. Items chosen as depending less on the language factor were scored as a possible non-verbal scale of intelligence, but the small number of available items has mitigated against respectable reliability and validity. Perhaps if the non-verbal items were augmented by other similar items, an adequate non-verbal scale could be constructed. The 'memory' scale is not reliable enough for diagnostic use, and even if it were, one might raise the question as to whether 'memory' rather than general intelligence is being tapped. After all, the so-called memory items were originally

selected, not because they were memory items but because they satisfied certain criteria as measures of general intelligence. Those who need scales for measuring 'pure' memory and other special abilities should look elsewhere.

Units of Measurement. - Chapter XI was devoted to the problem of units of measurement and to a discussion of some of the objections raised by one of the critics of age scales. As regards units of measurement, we have presented an argument which, it seems to us, completely nullifies the current notion that standard scores, or T-scores, are 'truly' equal units. Our examination of Kuhlmann's claim that the personal constant of Heinis is more constant than the I.Q. has led us to doubt the tenability of this claim.

Appendix A

NOTE ON SPURIOUS INDEX CORRELATION BETWEEN I.Q.'S

It has long been known that the correlation between two indices having a common variable denominator may be spurious, but in four recent statistical textbooks the 'may be' has been misconstrued as meaning 'is' or 'will be.' When sets of I.Q.'s from two tests are correlated, it is said by Garrett that the correlation will be spurious, by Guilford that such an 'r is spuriously high,' by Cooke that this is a 'source of spurious relationship,' while Peters and Van Voorhis cite the correlation of I.Q.'s, also E.Q.'s and A.Q.'s, as examples of spurious correlation.[1] It is the purpose of this note to show under what specific conditions such a correlation is or is not spurious.

We have approached this problem by two different methods, but since each leads to the same conclusions, we will here present the less elaborate approach. Indeed, the matter seems so elementary that we should be apprehensive about this very simple solution if its outcome had not been checked by more complex reasoning. It is self-evident that spuriousness could only exist in case age varied. The correlation between I.Q.'s for age constant cannot be spurious, so let us set up the correlation between two I.Q.'s, x and y, in the form of a partial with

[1]Cf. H. E. Garrett, *Statistics in Psychology and Education.* New York: Longmans Green Company, 1937, p. 458. J. P. Guilford, *Psychometric Methods,* New York: McGraw-Hill Book Company, 1936, p. 374. D. H. Cooke, *Minimum Essentials of Statistics,* New York: Macmillan Company, 1936, p. 187. C. C. Peters and W.R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases,* New York: McGraw-Hill Book Company, 1940, p. 217.

age, a, to be partialed out. Thus

$$r_{xy \cdot a} = \frac{r_{xy} - r_{xa}\, r_{ya}}{\sqrt{1 - r_{xa}^2}\ \sqrt{1 - r_{ya}^2}}$$

from which it is obvious that r_{xy}, the correlation between I.Q.'s, will be spuriously high if both sets of I.Q.'s are correlated in the same direction with age, spuriously low if these correlations are of opposite signs, and not spurious at all when the I.Q.'s are uncorrelated with age. Certainly the element of spuriousness is negligible for slight, e.g. ordinary chance sampling, departures of r_{xa} and r_{ya} from zero. For instance, suppose $r_{xa} = r_{ya} = .30$ (values this large will seldom arise as a result of sampling errors when N exceeds 100) and suppose r_{xy} is near .80, then the spuriousness will be less than .02.

An ideally constructed age scale will yield a correlation between I.Q. and age of zero for unselected cases. Although the New Revision meets this ideal, it must not be forgotten that groups can be so selected as to produce a non-chance correlation between I.Q. and age; in particular, a single school-grade group will tend to yield a negative correlation for these variables. We agree with the conclusion reached by Jackson in a recent paper[1] to the effect that every case involving the correlation of I.Q.'s must be considered individually, but we do not share his general alarm about the statistical analysis of I.Q.'s being meaningless. It might also be remarked that one of his simplifying assumptions, namely that M.A. variability equals that for C.A., is not tenable, not even as an approximation. For both age and grade sampling, the former tends to be the greater. Evidence for this may be found in Chapter III.

[1]R.W.B. Jackson, "Some Pitfalls in the Statistical Analysis of Data Expressed in the Form of IQ Scores," *J. Educ. Psychol.*, 1940, 31, 677-685.

Appendix B

ADJUSTMENT OF I.Q.'S FOR ATYPICAL VARIABILITY AT CERTAIN AGES

One of the requisites for an entirely satisfactory age scale is that the variability of the distributions of I.Q.'s be reasonably equal for successive ages. It has been admitted elsewhere (*Measuring Intelligence* , page 40) that the differences in variability for the New Revision are likely non-chance, and in Chapter VIII of this volume a factor was mentioned which possibly explains the low variability in the vicinity of ages 5 to 6, and the high variability at age 12. There seems to be no obvious reason for the high variability at ages 2-1/2 and 15. Since these differences mean that I.Q.'s for individuals are somewhat lacking in comparability, we present herewith a table by which one can make an adjustment to the I.Q.'s earned by individuals at certain ages. At the top of Table 52 will be found C.A.'s in years and months. If, for example, an individual has an obtained I.Q. of 120 and is 2 years and 7 months of age, the adjusted value would be 117; if he were of age 6, it would be 125, etc. The corrections are small for I.Q.'s near the average, and are larger for I.Q.'s which deviate from the average.

Appendix B

TABLE 52

I.Q. ADJUSTMENTS FOR VARIABILITY DIFFERENCES

Obtained I.Q.'s	2-4 to 3-3	4-10 to 6-6	11-6 to 12-5	14-6 to 15-5
148	140	159	142	143
146	139	157	140	141
144	137	154	138	139
142	135	152	137	137
140	134	149	135	136
138	132	147	133	134
136	130	144	131	132
134	129	142	130	130
132	127	139	128	128
130	125	137	126	127
128	124	134	124	125
126	122	132	123	123
124	120	130	121	121
122	119	127	119	120
120	117	125	117	118
118	115	122	116	116
116	113	120	114	114
114	112	117	112	112
112	110	115	110	111
110	108	112	109	109
108	107	110	107	107
106	105	107	105	105
104	103	105	103	104
102	102	102	102	102
100	100	100	100	100

TABLE 52. (Cont.)

I.Q. ADJUSTMENTS FOR VARIABILITY DIFFERENCES

Obtained I.Q.'s	2-4 to 3-3	4-10 to 6-6	11-6 to 12-5	14-6 to 15-5
98	98	98	98	98
96	97	95	97	96
94	95	93	95	95
92	93	90	93	93
90	92	88	91	91
88	90	85	90	89
86	88	83	88	88
84	87	80	86	86
82	85	78	84	84
80	83	75	83	82
78	81	73	81	80
76	80	70	79	79
74	78	68	77	77
72	76	66	76	75
70	75	63	74	73
68	73	61	72	72
66	71	58	70	70
64	70	56	69	68
62	68	53	67	66
60	66	51	65	64
58	65	48	63	63
56	63	46	62	61
54	61	43	60	59
52	60	41	58	57
50	58	39	56	56

Appendix C

ITEM CORRELATIONS WITH TOTAL SCORE

The retention of an item for the final forms depended in part upon its correlation with the total score based upon all the items of Forms L and M. We are giving in the following tables the biserial correlations for each item as computed on the age corresponding to its age placement. Actually, for many of the items, correlations were also determined at ages adjacent to the respective placement ages. In some cases, e.g. vocabulary, product-moment r's were also computed. Although all the available correlations were taken into consideration when retaining or eliminating items, we are presenting only one for each item. This should be sufficient to give the reader some notion of the degree of relationship between items and the total score.

The tables given herewith will also serve as a reference for identifying tests by age location and number, and name. More detailed descriptions of the tests can, of course, be found in *Measuring Intelligence* .

Appendix C

TABLE 53

FORM L BISERIAL CORRELATIONS: ITEMS VERSUS TOTAL SCORE

Location	Name of test	Scoring standard	r
L,II,1*	Three-hole form board	1	.67
2	Identifying objects by name	4	.64
3	Identifying parts of the body	3	.74
4	Block building: tower	±	.43
5	Picture vocabulary	2	.69
6*	Word combinations	±	.71
alt.	Obeying simple commands	2	.57
L,II-6,1	Identifying objects by use	3	.69
2	Identifying parts of the body	4	.43
3	Naming objects	4	.41
4	Picture vocabulary	9	.78
5	Repeating 2 digits	1	.62
6	Three-hole form board: rotated	1	.46
alt.	Identifying objects by name	5	.77
L,III,1	Stringing beads	4	.41
2	Picture vocabulary	12	.77
3*	Block building: bridge	±	.53
4	Picture memories	1	.61
5	Copying a circle	1	.35
6	Repeating 3 digits	1	.72
alt.*	Three-hole form board: rotated	2	.45
L,III-6,1	Obeying simple commands	3	.83
2	Picture vocabulary	15	.68
3	Comparison of sticks	3;5	.81
4	Response to pictures I	2	.50
5	Identifying objects by use	5	.84
6	Comprehension I	1	.70
alt.	Drawing designs: cross	±	.43

*Indicates duplicate test

Appendix C

TABLE 53 (Cont.)

FORM L BISERIAL CORRELATIONS

Location	Name of test	Scoring standard	r
L,IV,1	Picture vocabulary	16	.74
2	Naming objects from memory	2	.64
3	Picture completion: man	1	.47
4	Pictorial identification	3	.56
5	Discrimination of forms	8	.72
6	Comprehension II	2	.78
alt.	Memory for sentences I	1	.52
L,IV-6,1	Aesthetic comparison	3	.64
2	Repeating 4 digits	1	.58
3	Pictorial likenesses and differences	3	.62
4	Materials	2	.69
5	Three commissions	±	.51
6	Opposite analogies I	2	.78
alt.	Pictorial identification	4	.66
L,V,1	Picture completion: man	2	.48
2	Paper folding: triangle	±	.43
3	Definitions	2	.75
4	Copying a square	1	.46
5	Memory for sentences II	1	.70
6	Counting four objects	2	.57
alt.*	Knot	±	.49
L,VI,1	Vocabulary	5	.65
2	Copying a bead chain from memory I	±	.58
3	Mutilated pictures	4	.46
4	Number concepts	3	.56
5	Pictorial likenesses and differences	5	.73
6	Maze tracing	2	.60

Appendix C

TABLE 53 (Cont.)

FORM L BISERIAL CORRELATIONS

Location	Name of test	Scoring standard	r
L,VII,1	Picture absurdities I	3	.49
2	Similarities: two things	2	.71
3	Copying a diamond	2	.65
4	Comprehension III	2	.69
5	Opposite analogies I	5	.65
6	Repeating 5 digits	1	.60
L,VIII,1	Vocabulary	8	.75
2	Memory for stories: the wet fall	5	.70
3	Verbal absurdities I	3	.70
4	Similarities and differences	3	.75
5	Comprehension IV	2	.66
6	Memory for sentences III	1	.71
L,IX,1	Paper cutting I	1	.71
2	Verbal absurdities II	3	.73
3	Memory for designs	1	.28
4	Rhymes: new form	3	.52
5	Making change	2	.62
6	Repeating 4 digits reversed	1	.58
L,X,1	Vocabulary	11	.84
2	Picture absurdities II	±	.38
3	Reading and report	±	.49
4	Finding reasons I	2	.53
5	Word naming	28	.46
6	Repeating 6 digits	1	.45

Appendix C

TABLE 53 (Cont.)

FORM L BISERIAL CORRELATIONS

Location	Name of test	Scoring standard	r
L,XI,1	Memory for designs	1½	.46
2	Verbal absurdities III	2	.73
3*	Abstract words I	3	.89
4*	Memory for sentences IV	1	.53
5	Problem situation	±	.50
6*	Similarities: three things	3	.61
L,XII,1	Vocabulary	14	.79
2	Verbal absurdities II	4	.64
3*	Response to pictures II	±	.55
4	Repeating 5 digits reversed	1	.50
5	Abstract words II	2	.74
6	Minkus completion	2	.61
L,XIII,1	Plan of search	±	.57
2	Memory for words	1	.69
3	Paper cutting I	2	.56
4*	Problems of fact	2	.33
5*	Dissected sentences	2	.66
6	Copying a bead chain from memory II	±	.59
L,XIV,1	Vocabulary	16	.89
2	Induction	±	.75
3*	Picture absurdities III	±	.60
4	Ingenuity	1	.57
5	Orientation: direction I	3	.57
6	Abstract words II	3	.83

*Indicates duplicate test.

Appendix C

TABLE 53 (Cont.)

FORM L BISERIAL CORRELATIONS

Location	Name of test	Scoring standard	r
L,A.A.,1	Vocabulary	20	.78
2	Codes	1½	.48
3	Differences between abstract words	2	.80
4	Arithmetical reasoning	2	.64
5	Proverbs I	2	.75
6	Ingenuity	2	.68
7	Memory for sentences V	1	.72
8	Reconciliation of opposites	3	.56
L,S.A.I,1	Vocabulary	23	.76
2	Enclosed box problem	3	.42
3	Minkus completion	3	.51
4*	Repeating 6 digits reversed	1	.55
5	Sentence building	2	.67
6*	Essential similarities	2	.66
L,S.A.II,1	Vocabulary	26	.68
2	Finding reasons II	2	.37
3	Repeating 8 digits	1	.36
4	Proverbs II	2	.74
5	Reconciliation of opposites	5	.50
6*	Repeating thought of passage: value of life	±	.71
L,S.A.III,1	Vocabulary	30	.71
2	Orientation: direction II	2	.46
3	Opposite analogies II	2	.71
4	Paper cutting II	±	.52
5	Reasoning	±	.62
6	Repeating 9 digits	1	.53

*Indicates duplicate test.

Appendix C

TABLE 54

FORM M BISERIAL CORRELATIONS: ITEMS VERSUS TOTAL SCORE

Location	Name of test	Scoring standard	r
M,II,1	Delayed response	2	.43
2	Identifying objects by name	4	.70
3	Identifying parts of the body	3	.75
4*	Three-hole form board	1	.67
5	Picture vocabulary	2	.70
6*	Word combinations	±	.71
alt.	Naming objects	3	.49
M,II-6,1	Identifying objects by use	4	.77
2	Motor coordination	±	.46
3	Naming objects	4	.65
4	Picture vocabulary	7	.77
5	Repeating 2 digits	1	.56
6	Obeying simple commands	2	.59
alt.	Stringing beads	2	.44
M,III,1*	Block building: bridge	±	.53
2	Picture vocabulary	10	.77
3	Identifying objects by use	5	.69
4	Drawing a vertical line	±	.46
5	Naming objects	5	.55
6	Repeating 3 digits	1	.71
alt.*	Three-hole form board: rotated	2	.45
M,III-6,1	Comparison of balls	3;5	.82
2	Patience: pictures	1	.54
3	Discrimination of animal pictures	4	.66
4	Response to pictures I (level D)	2	.60
5	Sorting buttons	±	.66
6	Comprehension I	1	.76
alt.	Matching objects	3	.74

*Indicates duplicate test.

Appendix C

TABLE 54 (Cont.)

FORM M BISERIAL CORRELATIONS

Location	Name of test	Scoring standard	r
M,IV,1	Picture vocabulary	12	.63
2	Stringing beads	7	.33
3	Opposite analogies I	2	.62
4	Pictorial identification	3	.49
5	Number concept of two	2	.69
6	Memory for sentences I	2	.68
alt.	Discrimination of animal pictures	6	.64
M,IV-6,1	Discrimination of animal pictures	7	.60
2	Definitions	2	.45
3	Repeating 4 digits	1	.80
4	Picture completion: bird	1	.67
5	Materials	2	.69
6	Comprehension II	1	.63
alt.	Patience: pictures	2	.67
M,V,1	Picture vocabulary	14	.76
2	Number concept of three	2	.61
3	Pictorial similarities and differences	9	.64
4	Patience: rectangles	2	.49
5	Comprehension II	2	.64
6	Mutilated pictures	3	.66
alt.*	Knot	±	.49
M,VI,1	Number concepts	3	.57
2	Copying a bead chain	±	.40
3	Differences	2	.52
4	Response to pictures I (level II)	2	.61
5	Counting 13 pennies	1	.49
6	Opposite analogies I	4	.68

*Indicates duplicate test.

TABLE 54 (Cont.)

FORM M BISERIAL CORRELATIONS

Location	Name of test	Scoring standard	r
M,VII,1	Giving the number of fingers	±	.65
2	Memory for sentences II	1	.62
3	Picture absurdities I	2	.65
4	Repeating 3 digits reversed	1	.70
5	Sentence building I	2	.61
6	Counting taps	3	.31
M,VIII,1	Comprehension III	2	.71
2	Similarities: two things	2	.83
3	Verbal absurdities I	3	.74
4	Naming the days of the week	±	.61
5	Problem situations	2	.50
6	Opposite analogies II	3	.72
M,IX,1	Memory for designs I	1	.54
2	Dissected sentences I	2	.74
3	Verbal absurdities II	3	.65
4	Similarities and differences	4	.68
5	Rhymes: old form	2	.64
6	Repeating 4 digits reversed	1	.44
M,X,1	Block counting	8	.34
2	Memory for stories I: the school concert	5	.66
3	Verbal absurdities III	2	.77
4	Abstract words I	2	.59
5	Word naming: animals	12	.38
6	Repeating 6 digits	1	.34

Appendix C

TABLE 54 (Cont.)

FORM M BISERIAL CORRELATIONS

Location	Name of test	Scoring standard	r
M,XI,1	Finding reasons	2	.51
2	Copying a bead chain from memory	±	.38
3	Verbal absurdities II	4	.84
4*	Abstract words II	3	.89
5*	Similarities: three things	3	.61
6*	Memory for sentences III	1	.53
M,XII,1	Memory for designs II	±	.56
2*	Response to pictures II	±	.55
3	Minkus completion	2	.49
4	Abstract words I	3	.81
5	Picture absurdities II	±	.66
6	Repeating 5 digits reversed	1	.47
M,XIII,1	Plan of search	±	.46
2	Memory for stories II: acrobat	5	.27
3*	Dissected sentences II	2	.66
4	Abstract words II	4	.91
5*	Problems of fact	2	.33
6	Memory for sentences IV	1	.48
M,XIV,1	Reasoning	±	.54
2*	Picture absurdities III	±	.60
3	Orientation: direction I	3	.60
4	Abstract words III	2	.83
5	Ingenuity	1	.56
6	Reconciliation of opposites	2	.61

*Indicates duplicate test.

TABLE 54 (Cont.)

FORM M BISERIAL CORRELATIONS

Location	Name of test	Scoring standard	r
M,A.A.,1	Abstract words III	4	.90
2	Ingenuity	2	.74
3	Opposite analogies III	1	.78
4	Codes I	1½	.53
5	Proverbs I	2	.75
6	Orientation: direction I	4	.75
7	Essential differences	2	.84
8	Binet paper cutting	±	.53
M,S.A.I,1	Minkus completion	3	.55
2	Opposite analogies IV	1	.40
3*	Essential similarities	2	.66
4*	Repeating 6 digits reversed	1	.55
5	Sentence building II	2	.86
6	Reconciliation of opposites	4	.60
M,S.A.II,1	Proverbs II	1	.61
2	Ingenuity	3	.43
3	Essential differences	3	.56
4	Repeating 8 digits	1	.47
5	Codes II	1	.55
6*	Repeating thought of passage I: value of life	±	.71
M,S.A.III,1	Proverbs III	2	.66
2	Memory for sentences V	±	.29
3	Orientation: direction II	2	.57
4	Repeating 9 digits	1	.42
5	Opposite analogies IV	2	.69
6	Repeating thought of passage II: tests	±	.77

*Indicates duplicate test.